Transcriptional regulation and steady-state modeling of metabolic networks

Zelezniak, Aleksej; Kielland-Brandt, Morten; Patil, Kiran Raosaheb

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Transcriptional regulation and steady-state modeling of metabolic networks

Aleksej Zelezniak

PhD Thesis

Center for Microbial Biotechnology
Department of Systems Biology
Technical University of Denmark
December, 2012

"I have a friend who's an artist and has sometimes taken a view which I don't agree with very well. He'll hold up a flower and say "look how beautiful it is," and I'll agree. Then he says "I as an artist can see how beautiful this is but you as a scientist take this all apart and it becomes a dull thing," and I think that he's kind of nutty. First of all, the beauty that he sees is available to other people and to me too, I believe…

I can appreciate the beauty of a flower. At the same time, I see much more about the flower than he sees. I could imagine the cells in there, the complicated actions inside, which also have a beauty. I mean it's not just beauty at this dimension, at one centimeter; there's also beauty at smaller dimensions, the inner structure, also the processes. The fact that the colors in the flower evolved in order to attract insects to pollinate it is interesting; it means that insects can see the color. It adds a question: does this aesthetic sense also exist in the lower forms? Why is it aesthetic? All kinds of interesting questions which the science knowledge only adds to the excitement, the mystery and the awe of a flower. It only adds. I don't understand how it subtracts."

–Richard Feynman, What Do You Care What Other People Think, 1988

Acknowledgements

Most importantly I am very grateful to my supervisor Kiran Patil. Without his enthusiasm and ideas this work would not be possible. Kiran always supported and helped me in the hardest and most impossible situations. I am also very thankful to him for the opportunity of visiting his group at EMBL in Heidelberg. Working at EMBL has opened my view to many different sides of the beautiful world of a science. I am thankful to my supervisor Prof. Morten Kielland-Brandt for providing me the opportunity to work on my PhD project at the Center for Microbial Biotechnology in Denmark. I thank both of my supervisors, DTU and Novozymes for financial support of this work.

I want to thank all the members of Patil group, colleagues, office mates and friends who made my life easier during my PhD time. It is difficult to rank your help and support, so I have chosen to represent you alphabetically: Ana Rita Brochado, Ana Paula Oliveira, Anne Nørrevang Johansen, Arnau Montagud, Bo Salomonsen, Claudio Alfieri, Joana Xavier, Kalliopi Trachana, Lars Poulsen, Maria Secrier, Melanie Schmid, Olga Ponomarova, Pablo Minguez, Sara Lieder, Sergej Andrejev, Takuji Yamada, Tomasz Boruta, Tomas Stucko, Veli Vural Uslu, Zita Soons. A special thank goes to Steven Sheridan for fighting with my broken English and explicitly commenting on this thesis.

I am very thankful to my family for their support, patience and understanding...

Summary

Biological systems are characterized by a high degree of complexity wherein the individual components (*e.g.* proteins) are inter-linked in a way that leads to emergent behaviors that are difficult to decipher. Uncovering system complexity requires, at least, answers to the following three questions: what are the components of the systems, how are the different components interconnected and how do these networks perform the functions that make the resulting system behavior? Modern analytical technologies allow us to unravel the constituents and interactions happening in a given system; however, the third question is the ultimate challenge for systems biology. The work of this thesis systematically addresses this question in the context of metabolic networks, which are arguably the most well characterized cellular networks in terms of their constituting components and interactions among them. Furthermore, there is large interest in understanding and manipulating cellular metabolism from health as well as biotechnological perspectives. Fundamentally different biological questions are investigated in different core chapters of the thesis, though all are linked by the common thread of the functioning of cellular metabolism. The three main topics addressed are: i) transcriptional regulation of metabolite concentration, ii) transcriptional dys-regulation of skeletal muscle metabolism in type 2 diabetes, and iii) metabolic interactions in microbial ecosystems. The overall objective is to obtain novel understanding underlying the operating principles of metabolic networks.

Cellular responses to environmental perturbations and genetic/epigenetic modifications are to a large extent controlled through transcription, which is one of the fundamental mechanism/means of cellular regulation. An important question is to what extent gene expression can explain metabolic phenotype, in other words, how well changes in metabolite concentrations can be explained by the changes in related enzyme-coding transcripts? Attempts to predict changes in the metabolome from gene expression data have so far remained unsolved. Here, I challenge this question by proposing a mechanistic explanation of the interplay between metabolite concentrations, transcripts and fluxes based on Michaelis-Menten kinetics at the network-scale. The work demonstrates that in steady-state systems, changes of intracellular metabolites concentrations are linked with the changes in gene expression of both reactions that produce and reactions that consume a given metabolite. Analysis of a large compendium of gene expression data further suggested that, contrary to previous thinking, transcriptional regulation at metabolic branch points is highly plastic and, in several cases, the objective of the regulation appears to be metabolite-oriented as opposed to pathway-oriented.

The study thus provides a fundamental and novel view of metabolic network regulation in *Saccharomyces cerevisiae*.

Metabolism is a conserved system across all domains of life. Nowadays, metabolism has become a focal point in diagnosing and treating diseases such as diabetes and cancer. Type 2 diabetes mellitus is a complex metabolic disease which is recognized as one of the largest threats to human health in the 21st century. Recent studies of gene expression levels in human tissue samples have indicated that multiple metabolic pathways are dys-regulated in diabetes and in individuals at risk for diabetes; which of these are primary, or central to disease pathogenesis, remains a key question. Cellular metabolic networks are highly interconnected and often tightly regulated; any perturbations at a single node can thus rapidly diffuse to the rest of the network. Such complexity presents a considerable challenge in pinpointing key molecular mechanisms and signatures associated with insulin resistance and type 2 diabetes. The present work addresses this problem by using a methodology that integrates gene expression data with the human cellular metabolic network. The approach is demonstrated by analysis of two skeletal muscle gene expression datasets. The proposed methodology identified transcription factors and metabolites that represent potential targets for therapeutic agents and future clinical diagnostics for type 2 diabetes and impaired glucose metabolism. In a broader context, the study provides a framework for analysis of gene expression datasets from complex heterogeneous diseases, genetic, and environmental perturbations that are reflected in and/or mediated through changes in metabolism.

In nature, microorganisms do not exist as pure cultures, but evolve and co-exist with other species. Microbial communities have a variety of potential applications, including metabolic disease therapies and biotechnology. For example, microbial consortia consisting of various bacteria and fungi are known to exhibit a biodegradation performance superior to pure cultures, making them attractive research targets. It is believed that nutrition plays a crucial role in shaping microbial communities. Interspecies metabolite cross-feeding can confer several advantages to the community as a whole. For example, more efficient and complete use of available nutrients, or increased ability to survive under diverse/changing nutrition availability potentially induces fitness of individuals. The third topic of this thesis investigates the role of metabolic interaction in co-occurring microbial communities. The study aims to identify metabolic properties that shape the community structures. The analysis based on a global metagenomic dataset and genome-scale metabolic models suggested that species within coexisting communities have higher potential of metabolic cooperation compared to random controls. This work yielded a novel methodology (termed species metabolic coupling analysis) for

studying metabolic interaction and interdependencies within microbial communities. Species metabolic coupling analysis has a spectrum of applications to real-world problems, including investigation of metabolic interactions within the human microbiome, host-pathogen interactions and development of stable microbial communities.

Overall, this work contributes with novel insights, tools and methodologies to study the operation of cellular metabolism.

Dansk Sammenfatning

Biologiske systemer er karakteriseret ved en høj grad af kompleksitet, hvori de individuelle komponenter (f.eks. proteiner) er indbyrdes forbundet på en måde, der fører til en opførsel, der er vanskelig at forstå i detaljer. Udredning af systemets kompleksitet kræver i det mindste svar på følgende tre spørgsmål: hvad er komponenterne af systemerne, hvordan er de forskellige komponenter sammenkoblet, og hvordan udfører disse netværk de funktioner, der resulterer i systemernes adfærd? Moderne analytiske teknologier giver os mulighed for at optrævle de bestanddele og interaktioner der findes i et givet system, men det tredje spørgsmål er den ultimative udfordring for systembiologi. Nærværende afhandling behandler dette spørgsmål systematisk i forbindelse med metaboliske netværk, som velsagtens er de mest velbeskrevne biologiske netværk hvad angår komponenter og samspillet mellem dem. Desuden er der stor interesse for at forstå og manipulere cellestofskiftet ud fra såvel sundhedsmæssige som bioteknologiske perspektiver. Fundamentalt forskellige biologiske spørgsmål undersøges i forskellige centrale kapitler i afhandlingen, selv om de alle er forbundet af det fælles tema omkring, hvordan det cellulære stofskifte fungerer. De tre vigtigste emner, der behandles, er: i) Transkriptionel regulering af metabolit-koncentrationer, ii) transkriptionel dys-regulering af skeletmuskulaturens stofskifte i type-2 diabetes, og iii) metaboliske interaktioner i mikrobielle økosystemer. Det overordnede mål er at opnå ny forståelse bag de operationelle principper for metaboliske netværk.

Cellers reaktioner på forstyrrelser i vækstvilkår og genetiske/epigenetiske ændringer styres i høj grad gennem transkription, som er en af de grundlæggende mekanismer for cellulær regulering. Et vigtigt spørgsmål er, i hvilket omfang genekspression kan forklare metaboliske fænotyper; med andre ord, hvor godt kan ændringer i metabolitkoncentrationer forklares med ændringer i mængderne af mRNA kodende for de ansvarlige enzymer? Forsøg på at forudsige ændringer i metabolomet ud fra genekspressionsdata har hidtil ikke ladet sig gøre. Her udfordrer jeg dette spørgsmål ved at foreslå en mekanistisk forklaring af samspillet mellem metabolitkoncentrationer, transkripter og flux baseret på Michaelis-Menten kinetik på netværks-skala. Dette arbejde viser, at i steady-state systemer er ændringer i intracellulære metabolit-koncentrationer forbundet med ændringer i genekspression af både reaktioner, der producerer, og reaktioner, der forbruger en bestemt metabolit. I modsætning til tidligere tænkning tyder analyse af en stor samling af genekspressionsdata endvidere på, at transkriptionel regulering ved metaboliske forgreningspunkter er meget plastisk, og i flere tilfælde synes den selektive fordel ved reguleringen at være metabolit-orienteret snarere end pathway-

orienteret. Undersøgelsen giver således et fundamentalt og nyt syn på metabolisk netværksregulering i *Saccharomyces cerevisiae*.

Metabolisme er et i høj grad bevaret system på tværs af hele biologien. I dag er stofskifte blevet et centralt punkt i diagnosticering og behandling af sygdomme såsom diabetes og kræft. Type 2-diabetes mellitus er en kompleks metabolisk sygdom, der er anerkendt som en af de største trusler mod menneskers sundhed i det 21. århundrede. Nylige undersøgelser af genekspressionsniveauer i humane vævsprøver har vist, at flere metaboliske veje er dysreguleret i diabetes og hos personer med risiko for diabetes; hvilke af disse veje der er primære og/eller centrale for patogenesen, er fortsat et centralt spørgsmål. Cellulære metaboliske netværk er meget tæt forbundne og ofte stramt regulerede; eventuelle forstyrrelser ved et enkelt forbindelsespunkt kan således hurtigt udbrede sig til resten af netværket. En sådan kompleksitet udgør en betydelig udfordring i at indkredse de vigtigste molekylære mekanismer og kendetegn, der er forbundet med insulinresistens og type 2 diabetes. Det foreliggende arbejde løser dette problem ved at bruge en metode, der integrerer genekspressionsdata med det humane cellulære metaboliske netværk. Denne fremgangsmåde demonstreres ved analyse af to datasæt fra skeletmusklers genekspression. Den foreslåede metode identificerede transkriptionsfaktorer og metabolitter, der udgør potentielle mål for farmaka og fremtidig klinisk diagnose for type 2-diabetes og forringet glukosemetabolisme. I en bredere sammenhæng frembyder undersøgelsen en ramme for analyse af genekspression-data indsamlet ved komplekse heterogene sygdomme, genetiske og miljømæssige perturbationer, der afspejles i og/eller er medieret via ændringer i stofskiftet.

I naturen eksisterer mikroorganismer normalt ikke som rene kulturer, men udvikler sig og sameksisterer med andre arter. Mikrobielle samfund har en bred vifte af mulige anvendelser, herunder behandling af metaboliske sygdomme og bioteknologi. Eksempelvis kan mikrobielle konsortier bestående af forskellige bakterier og svampe udføre biologisk nedbrydning bedre end rene kulturer, hvilket gør dem attraktive at udforske. Det er almindeligt antaget, at ernæring spiller en afgørende rolle i udformningen af mikrobielle samfund, og indbyrdes udveksling og udnyttelse af metabolitter kan give flere fordele for samfundet som helhed. For eksempel kan en mere effektiv og fuldstændig anvendelse af de tilgængelige næringsstoffer, eller en forbedret evne til at tilpasse sig skiftende ernæringsforhold, føre til forbedret overlevelse af individerne. Det tredje emne i denne afhandling undersøger de metaboliske interaktioners rolle i blandede mikrobielle samfund. Formålet med undersøgelsen er at identificere de egenskaber ved metabolismen, der er bestemmende for strukturerne af de blandede samfund. Analysen er baseret på et globalt metagenomisk datasæt, og

metaboliske modeller i genom-skala pegede på, at arter inden for sameksisterende samfund har et større potentiale for metabolisk samarbejde i forhold til tilfældigt sammensatte samfund. Dette arbejde førte til en ny metode (kaldet species metabolic coupling analysis) for at studere metabolisk interaktion og indbyrdes afhængighed inden for mikrobielle samfund. Metoden har en vifte af konkrete anvendelser, herunder undersøgelse af metaboliske interaktioner i menneskets mikrobiom, værtspatogene interaktioner og udvikling af stabile mikrobielle samfund.

Samlet set bidrager dette arbejde med nye indsigter, værktøjer og metoder til at studere hvordan cellulært stofskifte fungerer.

List of Abbreviations

CoCCoA	Concentration change coupling analysis
EHMN	Edinburgh human metabolic network
FBA	Flux balance analysis
FCA	Flux coupling analysis
FH+/-	Family history positive/negative
FVA	Flux variability analysis
IGT	Impaired glucose tolerance
LP	Linear programming
MIP	Metabolic interaction potential
MILP	Mixed integer linear programming
MM	Michaelis-Menten
MOP	Metabolic resource overlap
NGT	Normal glucose tolerance
ORF	Open reading frame
OTU	Operational taxonomic unit
OXPHOS	Oxidative phosphorylation
prod/cons	Production/consumption
SMETANA	Species metabolic coupling analysis
T2DM	Type 2 diabetes mellitus
TF	Transcription factor
TSS	Transcription start site
WHO	World health organization

Contents

List of Figures

Chapter 1

Introduction and work objectives

Modern analytical technologies generate larger and larger 'omics' data sets across all branches of life sciences. The revolution in large-scale data generation continues to drive advances in modern biology, leading to conceptual developments and novel discoveries. However, current technologies focus on generating data, or, in other words assessing the concentrations of components but not their functions in biological systems (Sauer & Zamboni, 2008). Biological systems are extremely complex by their nature. A prerequisite for attaining understanding of any system is the knowledge of its constituent components and, most important, the interactions between those components. One the greatest challenge of modern biology is to transform the massive amounts of generated data into the biological knowledge. Therefore, for understanding underlying biological phenomena there is a general need to develop new approaches and tools that are capable of using multi-level large scale biological data for extraction of relevant biological signal. In particular, it is essential to identify *linkages* and mechanisms which underlie the logic of cellular organization and bring together interactions and structural components making a functional core of a cellular system. Metabolism is the fundamental basis of life at all scales, from the enzymatic reactions happening in each cell all the way to the multicellular organisms and the ecosystem as a whole. The research of this thesis is focused on i) identification the principles behind the regulation and operation of cellular metabolic network in model eukaryote *Saccharomyces cerevisiae*, ii) application of systems level analysis to human metabolic network in context of perturbed networks, iii) and illustrate the critical role of metabolism in microbial ecosystems.

The study of biochemical reactions and metabolism has been at the center of biological research since the nineteenth century (Fraenkel, 2011). However, in past 30 years under the shadow of advances in the field of molecular biology (McKnight, 2010) some scientists treated metabolism as a field 'to be mastered and put aside' (Bar-Even et al, 2012; Ray, 2010). Nevertheless, in recent years, the scientific community witnessed a rebirth in metabolic research as its part has been recognized playing a crucial role in many real-world applications (Bar-Even et al, 2012; Heinemann & Sauer, 2010; Klitgord & Segre, 2011; Reaves & Rabinowitz, 2011) For example, metabolism research has become crucial in diagnosing and treating diseases such as type 2 diabetes and cancer (Muoio & Newgard, 2008; Vander Heiden, 2011). Also, much of today's research is focused on addressing emerging challenges in sustainable energy, environmental chemistry and pharmaceutical industry by "squeezing" metabolism of microbial cell factories in the most needed way (Bar-Even et al, 2012). To

study, manipulate and redesign a metabolic system, one has to have a solid understanding of the biochemical and regulatory principles governing it.

In contrast to other biological networks, the functions and structure of metabolism are well understood (Gerosa & Sauer, 2011). Analytical techniques for quantitative measurements of metabolism components (metabolites, enzymes, fluxes) are available (Sauer & Zamboni, 2008). The current progress in systems level understanding of cellular metabolism was recently reviewed by Sauer and colleagues (Gerosa & Sauer, 2011; Heinemann & Sauer, 2010). Despite all the recent advances of analytical techniques, what we are still lacking is the knowledge of principles of regulatory networks that control metabolism functioning. Generally, we are missing the true system understanding of metabolism regulation in the complex genotype-to-phenotype path. For example, often it is unknown how metabolite concentration emerges from the interactions in the underlying network (Gerosa & Sauer, 2011); in other words, it is not clear how cells control metabolite pools in response to changing nutrient availability. Therefore, one of the objectives of the present work was to investigate possible quantitative mechanisms underlying regulation of metabolite concentration. **Chapter 3** and **Chapter 4** describe fundamentally different approach looking at the metabolite regulation. The presented integrative method provides novel quantitative mechanistic insights of transcriptional regulation of *Saccharomyces cerevisiae* metabolism.

Detailed molecular data obtained from studies of biological systems also change our view on pathogenesis of human diseases (del Sol et al, 2010). Despite the current progress in disease gene discovery, we are lacking mechanistic understanding of cellular diseases. Even many heritable diseases cannot be explained by classic one-to-one genotype-phenotype models (Wang et al, 2011). A widely accepted theory (Barabasi et al, 2011; del Sol et al, 2010; Emmert-Streib & Dehmer, 2011) is that diseases result from perturbation of cellular networks. One of the best examples is type 2 diabetes (T2DM), which is considered as disease of 21st century having an enormous health and economic impact in many countries all over the world. Currently 347 million people worldwide diagnosed with diabetes, and this number is expected to increase by 60% by 2030 (WHO, 2012). This fact sparked my interest studying the metabolism of T2DM; in particular, I focused on elucidating metabolic regulation signatures in connection with T2DM to provide novel insights about the pathogenesis of disease and discover possible biomarkers of disease progression. **Chapter 5** demonstrates an example and potential of integrative methods for studying complex metabolic diseases. The proposed framework can be applied not only to finding molecular signatures in disease

related perturbations, but also to investigating the regulation of metabolic networks in microbes following genetic or environmental perturbations.

In nature, very few species exist in isolation; nearly all life forms are in some way interacting between each other. For example, microorganisms do not exist as pure isolated cultures, but usually form complex ecological interaction webs – ecosystems. Microorganisms have a large impact on our daily lives – they produce dairy products (Chen et al, 2009), are part of our immune system (Phelan et al, 2012), play important role in digestion (Arumugam et al, 2011), and clean our waste and pollution (Mikeskova et al, 2012). In equilibrium, microorganisms coexist in the stable communities and perturbation of these communities can have a catastrophic impact on our society, *e.g.* epidemics (Phelan et al, 2012). Interactions within these communities can have a positive impact, a negative impact or no effect on the individuals involved. All of these interactions, regardless their outcome happens through a diverse set of mechanisms by which molecular information (including DNA and proteins) is being exchanged (Phelan et al, 2012). Microbial communities are very diverse in their structures (Fuhrman, 2009), and it is not clear what types of interactions are most prevalent. There is a common believe that nutrition is a driving force that shapes microbial communities (Fuhrman, 2009). By cross-feeding each other, community as a group possibly could survive in more conditions and thus potentially induce fitness of individuals. However, due to technical limitations it is particularly challenging to probe interactions within a community, for example, determine which metabolites are being exchanged and what exactly is the role of these interactions in shaping communities. One way to tackle this problem is to model metabolic capabilities of individual members within a group of species. **Chapter 6** describes a method which was developed to model metabolic interactions between species of large size communities and provides an example of community metabolic modeling in co-occurring species derived from global environmental data.

Outline of thesis

The work presented in this thesis spans fundamentally different problems which are united by the common thread of functioning of metabolic networks with the overall objective of gaining novel understanding behind the operating principles of microbial metabolism. The thesis is organized in several chapters, starting from basic principles about the organization of cellular metabolism, and the following four chapters representing different biological applications. Each chapter introduces the reader to the biological problem, provides its own results, the discussion and methods sections and can be read independently. Readers familiar with the topic of metabolic networks may skip **0**;

however, before reading **Chapter 4** it is recommended that the reader have an understanding of the mechanistic principles described in **Chapter 3**. The work is concluded with summary and future directions (**Chapter 7**). An outline of the thesis is presented in Table 1.1

Table 1.1 Organization of the present work

General background	
Chapter 2	Introduction to metabolism and key properties of metabolic networks
Regulation of metabolic networks	
Chapter 3, Chapter 4	Novel mechanistic view about regulation of metabolite concentrations in *Saccharomyces cerevisiae*
Application of data integrative methods	
Chapter 5	An example of integrative systems analysis for studying metabolism related diseases, type 2 diabetes case study
Modeling metabolic interaction in microbial communities	
Chapter 6	New stoichiometric method for modeling metabolic interactions in microbial communities. Illustrates the role of metabolism in microbial coexistence
Conclusions and future perspectives	
Chapter 7	Summarizes the work and suggest the future directions in the analysis of biological systems

List of publications

The work described in my thesis resulted in the following publications. During my PhD project (November 2009 – October 2012) I also contributed to additional publications, which were not included in this thesis.

Articles included in the thesis:

Zelezniak A, Pers TH, Soares SP, Patti ME, Patil KR. Metabolic Network Topology Reveals Transcriptional Regulatory Signatures of Type 2 Diabetes. *PLoS Computational Biology*, Vol. 6, No. 4, 2010, p. e1000729

Zelezniak A, Sheridan S, Patil KR. Quantitative relationship between gene expression and metabolite levels is jointly determined by reaction mechanism and network connectivity. *Manuscript in preparation*

Zelezniak A, Andrejev S, Ponomarova O, Patil KR. Co-occurring bacterial communities feature high potential for metabolic cooperation. *Manuscript in preparation*

Articles not included in the thesis:

Yamada T, Waller AS., Raes J, **Zelezniak A,** Perchat N, Perret A, Salanoubat M, Patil KR, Weissenbach J, Bork P. Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours. *Molecular Systems Biology*, Vol. 8, 2012

Montagud A, **Zelezniak A**, Navarro E, de Córdoba PF; Urchueguía JF, Patil KR. Flux coupling and transcriptional regulation within the metabolic network of the photosynthetic bacterium *Synechocystis sp.* PCC6803. *Biotechnology Journal*, Vol. 6, No. 3, 2011, p. 330-342

Chapter 2
General background

Organization of cellular metabolism/metabolic networks

Metabolism can be described as a biological system which a cell acquires to fulfill its requirements by synthesizing the cellular components and energy needed for growth and maintenance. One of the most important challenges in modern biology is to understand the relationship between structure, function, and regulation of complex biological networks including metabolism. The biochemical transformations of metabolism are well studied and documented (Cherry et al, 2012; Joshi-Tope et al, 2003; Kanehisa, 2002; Kumar et al, 2012; Schellenberger et al, 2010) providing the basis for building large-scale, structurally accurate biological networks.

Conventionally, metabolism has been described and viewed as a set of discrete pathways. Pathways consist of sets of enzymes operating towards synthesis/degradation of certain metabolites or a group of metabolites. Although the concept of a pathway is widely used and is particularly useful for illustrative purposes (**Figure 2.1**), it has become increasingly clear that metabolism operates as a highly integrated network (Sweetlove et al, 2008). The definition of a pathway is not exact from the stoichiometric point of view. For example, synthesis of one metabolite often involves the operation of many pathways requiring energy and cofactors (*i.e.* NADH/CoA/ATP) which must be available to drive the reaction. The biochemical makeup of enzymes and metabolites varies between organisms, but in terms of stoichiometrically balanced reaction sets, the general picture of metabolism is similar across many organisms. The global overview of cellular metabolic pathways is depicted in **Figure 2.1**.

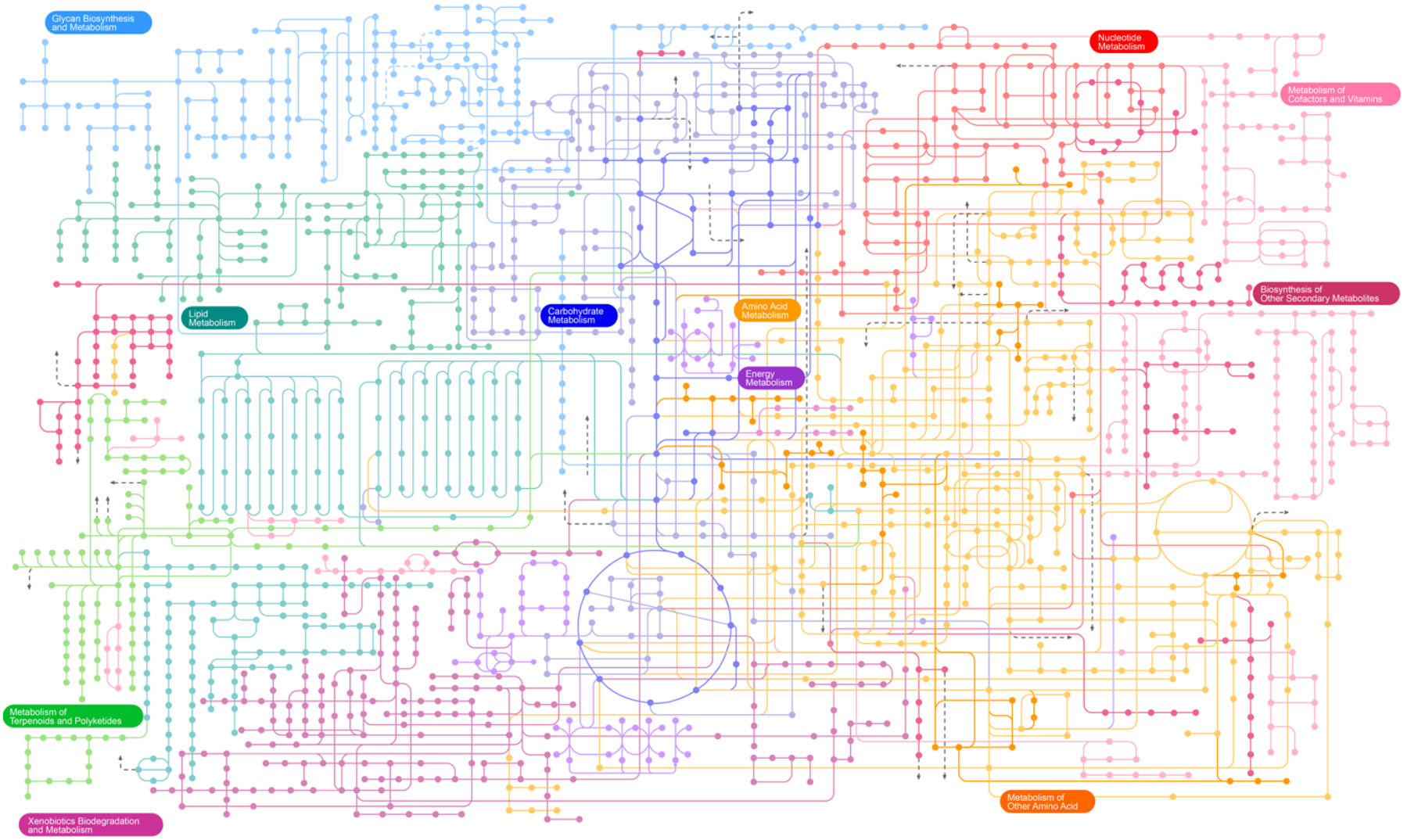


Figure 2.1 Global overview of metabolic pathways. Metabolism is arbitrarily divided into 11 parts according to the chemical type of compounds. Figure source: http://pathways.embl.de

In the post-genomic era, biological systems are viewed as networks which are often represented as graphs (Eisenberg et al, 2000). Particularly, for metabolic pathways, enzymes and metabolites are represented as nodes and interaction between them as edges. Edges can be directional or bidirectional. A metabolite node interacts with all of the enzyme nodes that catalyze a reaction involving that metabolite, and an enzyme node interacts with all of the metabolites which participate in the corresponding reactions. Each node is connected to a different type of node forming a directed bipartite graph (**Figure 2.2**), *e.g.* metabolite is only connected with enzyme, but two metabolites/enzymes cannot be connected together. Although such a representation visually is not the most intuitive, from the analytical point of view it provides a direct access to the properties of metabolic networks. For example, information such as in how many reactions ATP is being formed or which genes/enzymes regulate regulates this process, etc. can be easily accessed.

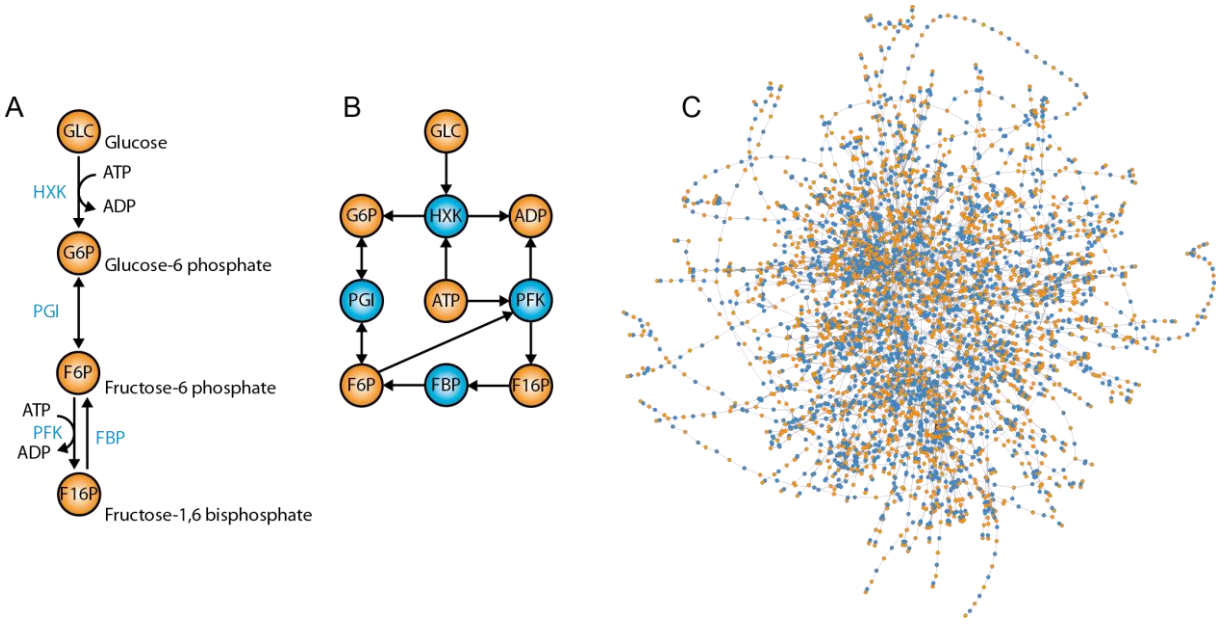


Figure 2.2 Different representations of metabolic pathways. A) Classical representation of metabolic reactions, highly connected metabolites such as ATP are depicted only in individual reactions. B) Same pathway represented as directed bipartite graph, where enzymes/metabolites are depicted as nodes and interaction as edges. C) Human metabolic reconstruction (Ma et al, 2007) represented as a bipartite graph. Due to metabolism complexity such network representation is not very clear for visual inspection.

Metabolic networks are structurally organized in such a way that a large variety of biochemical products and complex macromolecules (such as proteins, DNA, RNA) are synthesized from a variability of nutrients by conversion through relatively few common intermediate metabolites. The common intermediate metabolites and their enzymes form a hub or the knot of so-called bow-tie structure (Csete & Doyle, 2004). This bow-tie structure results in an interconnected core set of central reactions that constitute the backbone of high metabolic fluxes. For example, the central

carbon metabolism provides a variety of alternative pathways for generating essential precursor molecules, energy carrier molecules (ATP) and reduction equivalents (NAD(P)H) (Szallasi et al, 2010).

Metabolic networks usually are fully connected networks, meaning that it is possible to connect any of two nodes in the network (Patil, 2006). Another important property of metabolic networks as of many other biological networks is that they exhibit scale-free topology (Wolf et al, 2002). The concept of scale-free network (Szallasi et al, 2010) applies to the connectivity distribution of each node in the network. Compared with random networks, in which the number of connections is highest at the average value and decays exponentially, scale-free networks contain a relatively small number of highly connected nodes and distribution of a number of connections in nodes follows a power law (Barabasi & Oltvai, 2004). Such topology on a purely structural level exhibits the following main properties: i) '*small-world behavior*' that is any two nodes can be connected via a small number of intermediate nodes and ii) probably the most important property of scale-free networks is that they are robust in terms of high tolerance to errors. For example, random removal of a large fraction of nodes will not destroy a network, which can be observed in gene-knockout studies, deletion of certain enzymes (*e.g.* isoenzymes or enzymes involved in alternative pathways) does not affect the growth phenotype (Baba et al, 2006). Moreover, detailed studies of metabolic networks (Albert, 2005) showed that the power-law connectivity distribution is valid for organisms from all three domains of life – bacteria, archaea and eukarya. Metabolite-reaction connectivity distributions of several networks used in the present work are depicted in **Figure 2.3**.

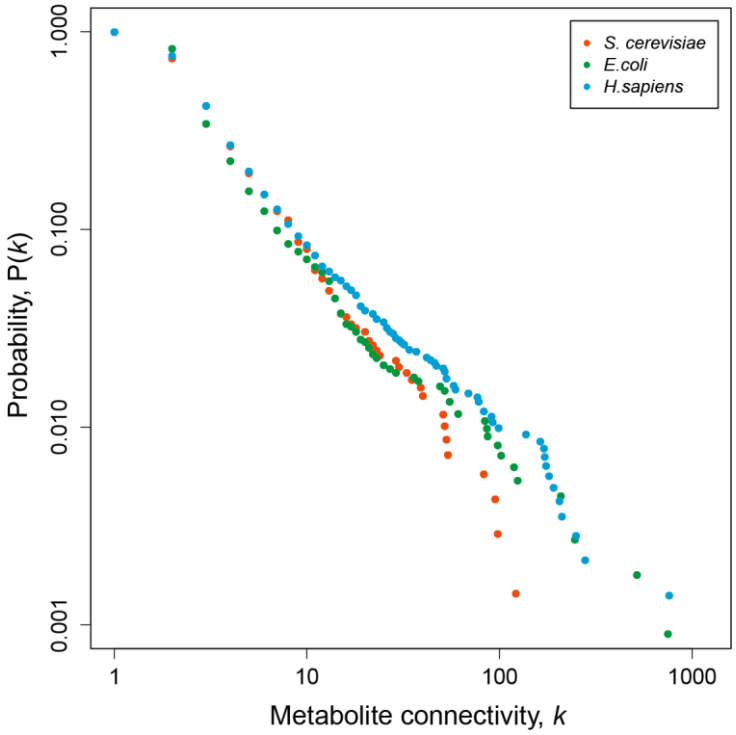


Figure 2.3 Metabolite connectivity distributions in metabolic networks. Metabolic networks are scale-free networks which are characterized by a power-law connectivity distribution. *k* represents a number of reactions where a given metabolite is involved, and probability P(*k*), denotes the probability that a selected metabolite has *k* connections (reactions). Both axes are in log 10 scale. Presented metabolic reconstructions (Duarte et al, 2007; Forster et al, 2003) were used in the present work. For *Escherichia coli*, automatic reconstruction from Model SEED pipeline (Henry et al, 2010) is shown.

Regulation of metabolic fluxes

As a result of evolution, all living forms developed a fine-tuned system to meet exactly demand for energy and building blocks needed for cellular growth and maintenance of biological processes (Nielsen, 2003). Before starting to discuss the concepts of metabolic regulation, it is necessary to formalize the term “metabolic flux”. Metabolic flux is the amount of substrate utilized (or products produced) by a reaction per unit time and usually it is normalized to some measurable parameter (for example, gram glucose utilized per hour per gram of cell mass) (Patil, 2006). When cell experiences environmental perturbation (changes of nutrients, pH, temperature, osmotic pressure, etc), it adjusts its metabolic network to a new steady-state to fit the new environment and this to a large extend happens through the sensing of metabolite concentrations. The cell reacts to the changes of metabolite concentrations, and, as a result, it sets the system to the new stable steady-state in order to keep metabolite pools at homeostasis (**Figure 2.4**). The concentration of a metabolite depends on its rate of degradation as well as its rate of synthesis. Moreover, metabolites are often synthesized simultaneously rather than in isolation from each other (Stitt et al, 2010).

Obviously, such regulation involves a large number of genes, and modifying the activity of the gene products (enzymes) influences the fluxes and consequently metabolite concentrations in the metabolic network. Metabolites and fluxes are the objects of metabolic regulation and enzymes are the tools by which the cell adjusts its metabolic phenotype. Exact biochemical mechanisms that link the metabolic network to transcription and translation are not well understood. Adjustments happen through the complex pathway of interactions from the sensor molecule action (through specific proteins or directly) on genetic information coding sequences (DNA or RNA), down to the functional proteins (enzymes) which in turn control metabolic reactions and consequently change metabolic phenotype of the cell.

The fluxes are constrained by the laws of thermodynamics. From the classical Michaelis-Menten (Briggs & Haldane, 1925; Michaelis & Menten, 1913) point of view reaction rates/fluxes are dependent on three biological quantities interacting at the level of enzyme kinetics: i) presence of active enzyme, ii) enzyme properties such as affinity for a substrate or presence of inhibitor and iii) metabolite concentrations. The concentration of the metabolite is itself a function of the flux and the properties of the enzymes; consequently, there is an enforced/continuous regulation in the system .

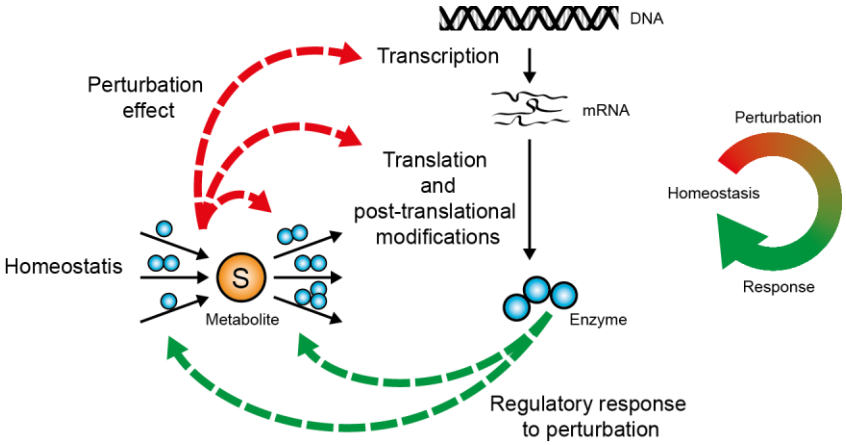


Figure 2.4 Metabolism regulation, from gene expression to metabolite levels. A cell experiencing perturbation "senses" metabolite concentration changes, affecting enzyme regulation via potential interactions with proteins and/or mRNA. As a regulatory response, metabolite levels are indirectly adjusted to the new stable state (homeostasis) at the genetic level, through enzyme abundances.

Modeling behavior of metabolic networks

Current sequencing technology is enabling production of thousands of sequenced genomes each year which facilitates reconstructions of genome-scale metabolic networks. Over the last few decades, there has been developed standards for reconstruction of metabolic networks (Feist et al, 2009; Thiele & Palsson, 2010), providing the basis for development of automatic metabolic

reconstruction platforms (Henry et al, 2010; Karp et al, 2002). Applications, tools and progress of use of metabolic network reconstructions have been recently reviewed (Feist et al, 2009; Oberhardt et al, 2009).

Apart from structural information, metabolic network reconstructions can be used to model intracellular fluxes. A complete mechanistic dynamic model of metabolism would provide the most detailed understanding of a biological system and would allow the dynamic prediction of cellular behavior. However, measurement of fluxes and metabolite concentrations at the entire metabolic network scale is still a difficult task owing to a variety of technological and experimental limitations. Moreover, such models require knowledge of reaction mechanisms and *in vivo* enzyme kinetic parameters which are very difficult to obtain making such models applicable only for certain parts of the cellular system. On the other hand, a combination of static models with dynamic components has been recently successfully applied to model cell cycle of *Mycoplasma genitalium,* providing a outstanding example of understanding dynamic cellular behavior (Karr et al, 2012).

Another modeling approach that gained community acceptance and popularity over the last decade is Flux Balance Analysis (FBA) (Varma & Palsson, 1994). A particular success of the method is that it is capable of quantitatively modeling carbon and energy metabolism for the genome-scale networks. FBA has been expanded for the variety of applications including finding the best targets for metabolic engineering strategies (Brochado et al, 2010; Burgard et al, 2003), large scale integrative modeling by combining different levels of information (Karr et al, 2012; Yizhak et al, 2010), microbial community models (Freilich et al, 2011; Klitgord & Segre, 2010; Stolyar et al, 2007; Zomorrodi & Maranas, 2012) and many more (Burgard et al, 2004; Lewis et al, 2012; Segre et al, 2002).

Briefly, the fundamental principle behind FBA relies based on conservation of mass. For each metabolite (X_i) in the metabolic network a *flux balance* can be written as:

$$\frac{dX_i}{dt} = V_{synthesis} - V_{\mathrm{degradation}} - (V_{biomass} - V_{extracellular})$$

Where V represents fluxes through which the metabolite X_i is synthesized, degraded, used for biomass composition and up-taken from or secreted to the extracellular environment. The biomass composition and extracellular metabolite uptake/secretion rates are determined experimentally. Inside the cell, biochemical transformations happen very rapidly compared to the cellular growth and uptake/secretion reactions. Thus the main assumption implies that the reactions are in the (quasi-) steady-steady, over a period of time their concentrations remain constant. This makes the fluxes

balanced, in other words the formation of metabolite must be balanced by the metabolite degradation fluxes in order to not accumulate the intracellular metabolite pool. Typically, the solutions to such problems are obtained through the formulation of a linear optimization problem, which optimizes a chosen objective function, *e.g.* the cellular growth. The biological objective of the cell has been always under debate (Schuetz et al, 2007; Schuetz et al, 2012); nevertheless, in laboratory conditions, the objective of growth has been shown to be in agreement with experimental data for some microbial species (Forster et al, 2003; Varma & Palsson, 1994).

One spectacular application of linear programming applied for metabolic networks is flux coupling analysis (Burgard et al, 2004). The methodology allows studying the topology of metabolism operation under physiological constraints. Flux coupling analysis is useful method for establishing reaction dependencies under given growth conditions.

Besides the constraint-based methods, another important class of useful techniques for analysis of properties and operational principles of metabolic networks is the so-called network-based pathway analysis, including elementary modes (Schuster et al, 1999) and extreme pathways (Schilling et al, 2000). The concept of network-based pathway analysis provides a computational tool describing all possible metabolic routes that could operate at steady-state (Schuster et al, 2000). Recent algorithmic improvements allow computing elementary flux modes for moderate size networks (Terzer & Stelling, 2008). However, the analysis of genome-scale metabolic networks still remains computationally challenging, due to the number of possible metabolic operation modes which grows exponentially with increasing network size and connectivity. In this thesis, metabolic networks are analyzed using constraint-based modeling approaches, and the current work does not involve network-based pathway analysis. Interested readers can find more information about network-based analytical tools in a recently published review (Trinh et al, 2009)

IT'S WEIRD HOW PROUD PEOPLE ARE OF NOT LEARNING MATH WHEN THE SAME ARGUMENTS APPLY TO LEARNING TO PLAY MUSIC, COOK, OR SPEAK A FOREIGN LANGUAGE.

[*] Figure source: http://imgs.xkcd.com/comics/forgot_algebra.png

Chapter 3

Quantitative relationship between gene expression and metabolite levels is jointly determined by reaction mechanism and network connectivity

Abstract*

Metabolite levels and their turnover rates are the main determinants of cellular metabolic state, which can only be indirectly regulated by the cell. Transcription is the first step towards regulating metabolism and elucidation of its link to metabolite concentrations is a fundamental challenge in metabolic systems biology. At the level of individual reactions, intra-cellular metabolite levels are dependent on enzyme abundances through reaction kinetics mechanisms. Another mechanism of metabolite concentration control is due to network connectivity, which has remained relatively less elucidated. We hereby integrate these two distinct mechanisms and develop a modeling framework describing the interplay between metabolite concentrations, mRNA levels and fluxes. We illustrate the key role of network connectivity in determining quantitative relationship between gene expression and metabolite levels through the analysis of experimental data from *Saccharomyces cerevisiae*. The proposed methodology provides mechanistic explanation for the network-guided organization of transcriptional response in metabolic networks and represents a step towards unraveling the genotype-phenotype relationship.

* Manuscript in preparation: Zelezniak A, Sheridan S, Patil KR. Quantitative relationship between gene expression and metabolite levels is jointly determined by reaction mechanism and network connectivity

Introduction

Regulation of intra-cellular metabolite levels in response to environmental and genetic changes is fundamental to the survival and evolution of organisms. Metabolism is one of the key cellular functionality at the interface of the genome and the organism's environment. On the cell side, metabolism provides basic building blocks and energy for the growth and maintenance. On the environment side, metabolism exchanges material and energy with the surroundings to maintain thermodynamic feasibility of cellular processes. Genetically encoded messages thereby undertake a long journey, through complex multi-level interaction layers, down to functional enzymes, in order to regulate and drive life's essential metabolic reactions. Emerging genome-wide bio-molecular abundance and interaction studies are providing snapshots of these regulatory networks (Costenoble et al, 2011; Fendt et al, 2010; Gallego et al, 2010; Li et al, 2010; Oliveira et al, 2008; Wang et al, 2010). It has been clear that the adjustments in cellular metabolic phenotype (*i.e.*, rates of reactions (fluxes) and metabolite levels) involve large numbers of changes at the gene expression levels (Murray et al, 2007; Tai et al, 2005; Tu et al, 2005). For example, previous studies have shown that the transcriptional regulation within metabolic networks is organized around perturbation-specific metabolites crucial for adjusting the network state (Patil & Nielsen, 2005; Zelezniak et al, 2010). Despite successful outcomes of these and other studies in linking gene expression with metabolites on a qualitative basis (Bradley et al, 2009; Keurentjes et al, 2006; Murray et al, 2007; Patil & Nielsen, 2005; Urbanczyk-Wochniak et al, 2003), the quantitative relationship between the transcriptional response and the corresponding changes in metabolite levels has remained largely elusive. The task of predicting metabolite level changes based on gene expression is challenging due to multiple layers of regulation involved in-between (**Figure 3.1**A).

We postulate that two primary mechanisms will largely determine the association between mRNA and metabolites, *viz.* reaction kinetics (which is non-linear by nature (Briggs & Haldane, 1925; Michaelis & Menten, 1913; Van Slyke & Cullen, 1914)) and mass flow constraints imposed by the network connectivity, *i.e.*, balance between the production and the consumption of a metabolite. Role of reaction kinetics has been previously examined, albeit in the context of isolated reaction-metabolite pairs. With such an approach, changes in metabolite levels could be explained to some extent when using protein abundance as a measure of enzyme availability; however, no correlation was observed in case of gene expression (Fendt et al, 2010). Detailed kinetic models and metabolic control analysis are useful tools to this end (Cleland, 1989; Klipp et al, 2005; Westerhoff & Chen, 1984), but are currently limited by the lack of availability of *in vivo* kinetic parameters. Consequently,

detailed kinetic simulations are unsuitable for many biological studies where a comparison between two states/factors is the focus (*e.g.* healthy *vs.* disease state).

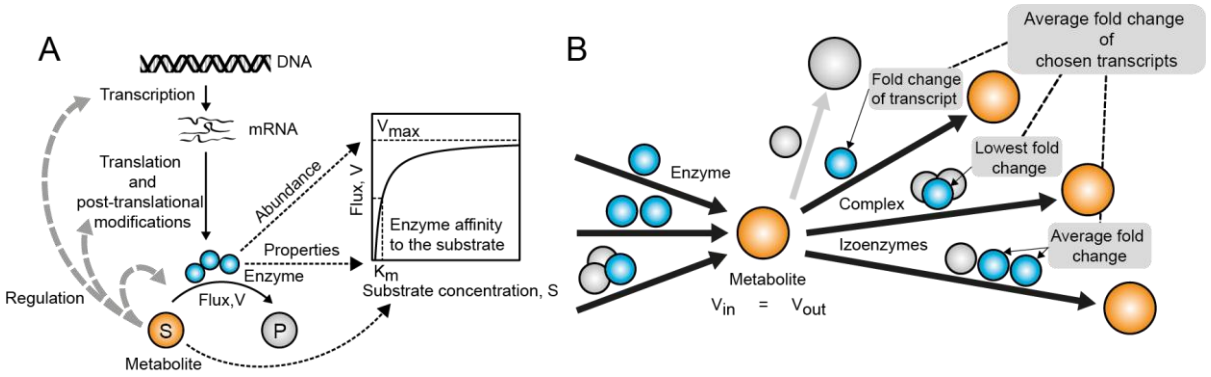


Figure 3.1 From gene expression to metabolite levels. A) Metabolite levels are indirectly regulated at the genetic level, through enzyme abundance. Structural properties of an enzyme determine affinity for its substrate (K_M), while the enzyme abundance limits the maximum substrate turnover rate (V_{max}) achievable. Substrate concentration is linked with the reaction rate by non-linear MM kinetics, parameterized by K_M and V_{max}. Metabolites, in turn, can feedback to the enzyme regulation via potential interactions with proteins (Gallego et al, 2010) and/or mRNA (Hentze & Preiss, 2010). B) Transcript-metabolite relationships are usually many-to-one. Either a single enzyme, or isoenzymes or a protein complex governs each reaction producing or consuming a metabolite. In our analysis, we discarded transcripts with insignificant changes ($\alpha = 0.05$) across conditions (grey circles). For the remaining transcripts we combined the corresponding fold changes to derive gene-expression scores for reactions and thereafter for consumption or production of metabolites.

In vivo, most metabolites participate in multiple reactions and thus the abundance of enzymes catalyzing individual reactions will not completely determine metabolite's fate, neither in terms of concentration nor fluxes. Indeed, correlations between mRNA and fluxes and even between enzyme activities and fluxes have often been found to be poor (Daran-Lapujade et al, 2007; Rossell et al, 2005; Rossell et al, 2006; Yang et al, 2002a). On the other hand, studies utilizing network connectivity have been successful in establishing gene-expression-metabolite relationship, however, only of qualitative nature (Bradley et al, 2009; Hirai et al, 2005; Kresnowati et al, 2006; Murray et al, 2007; Urbanczyk-Wochniak et al, 2003). Here we propose an integrative model that couples both mechanisms of concentration control, *i.e.*, reaction kinetics and network topology. In essence, the model embeds Michaelis-Menten kinetics into the metabolic network through the use of mass balance constraints. The model allowed us to link response at the level of metabolite concentration to the transcriptional fold changes in the neighboring as well as topologically distant genes. We verified the resulting model by analyzing publically available datasets reporting transcriptional and metabolic changes in the central carbon metabolism of the yeast *Saccharomyces cerevisiae*.

Results

Starting from the classical Michaelis-Menten (MM) model – looking at each reaction as an isolated system consisting of a single enzyme and its substrate, we develop a network kinetics approach by accounting for the interactions between different reactions through shared metabolites. By analogy to flux coupling analysis, which describes how steady-state fluxes are linked to each other (Forster et al, 2003), we term our approach Concentration Change Coupling Analysis (CoCCoA). In view of the lack of genome-wide detailed information on *in vivo* enzyme mechanisms, we consider a single-substrate MM kinetics for all reactions. According to MM kinetics, the flux or reaction rate V is a function of three parameters: i) concentration of substrate, S; ii) activity of enzyme, V_{max}; and iii) enzyme properties, K_M (**Figure 3.1**A). MM kinetics describes one enzyme – one substrate interaction. However, not only most intra-cellular metabolites participate in multiple reactions, several reactions are governed by multiple proteins (Forster et al, 2003). To account for these dependencies, reactions were classified in to three types: i) reactions controlled by a single enzyme, ii) reaction catalyzed by two or more isoenzymes, and iii) reactions catalyzed by enzyme complexes. For each reaction, we thereby chose the enzyme coding transcripts accordingly – in case of isoenzymes we average fold changes of transcripts, while for complexes we picked transcript with the lowest fold change. Applying these criteria, each reaction can be assigned a single fold-change value (in our analysis we used only significantly changed transcripts, $\alpha = 0.05$) (**Figure 3.1**B). Finally, for each metabolite we calculate a gene-expression based concentration response score according to one of the CoCCoA equations, which are developed in the following. The overall workflow used for the analysis is depicted in **Figure 3.2**. Genome-scale metabolic reconstruction of *S. cerevisiae* (Forster et al, 2003) was used as a basis for obtaining the network connectivity as well as reaction directionality information. We used three different experimental datasets for evaluating the proposed CoCCoA models. For each dataset, we compared the experimentally observed metabolite concentration changes with the CoCCoA scores as calculated by using gene expression data. These case studies include one genetic and two environmental perturbations where both gene expression and metabolite concentrations had been measured in the same experiment (Methods). Significance of the observed correlations was evaluated against correlations obtained by randomly permuting metabolite and CoCCoA scores (Methods).

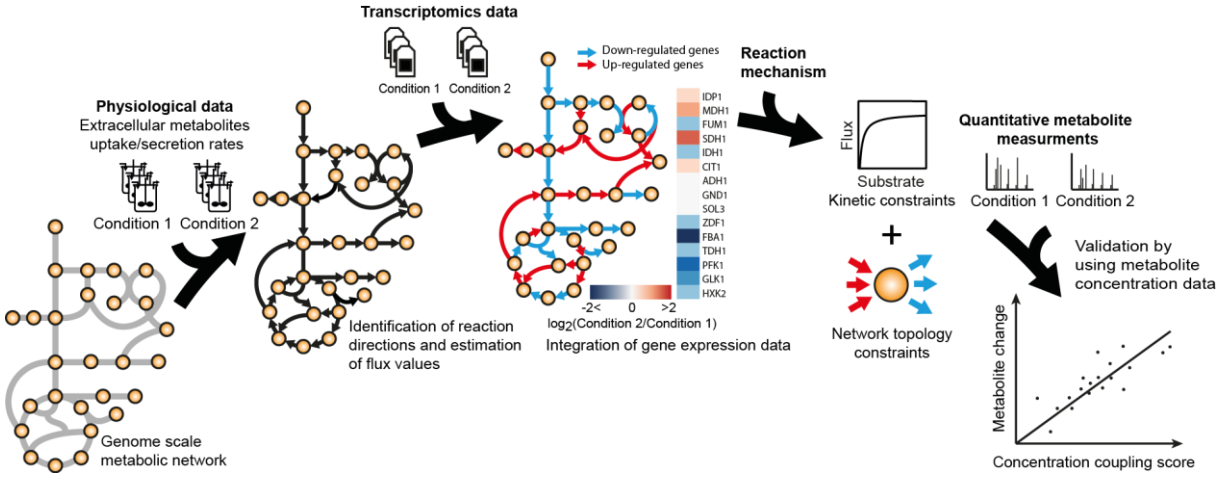


Figure 3.2 Schematic workflow of the algorithm used for the proposed concentration change coupling analysis. In the first step, physiological measurements from growth experiments are used to constrain the genome-scale metabolic model (Methods). Subsequent computational flux analysis (Methods) helps in identifying the directionality and range of fluxes under the two conditions being compared. Next, by using the comparative transcriptome data, fold changes at the individual gene-expression level are mapped on to the reactions in the network (**Figure 3.1**B). Concentration change coupling analysis integrates the mapped gene-expression data with the network topology by using a model formulation derived from reaction kinetics mechanism and mass balance constraints (main text). The main output from the algorithm is a measure of metabolite concentration changes between the two conditions, which are here tested for correlation with the experimentally measured metabolite abundance data.

Depending on the extent to which the network connectivity information is included, the CoCCoA models are termed 0^{th}, 1^{st} or 2^{nd} degree (**Figure 3.3**). 0^{th} degree CoCCoA considers only the consumption of a metabolite and thus relying on enzyme kinetics. 1^{st} degree CoCCoA accounts for the production of a metabolite in addition to its consumption by using mass balance constraints. Further expanding the scope of network connectivity accounted for in the model, the 2^{nd} degree CoCCoA includes producing reactions of pre-cursors of the metabolite in question. The expansion of the network information on the production side of metabolite (and not consumption) is justified through mass balance and enzyme kinetic constraints as outlined in the following.

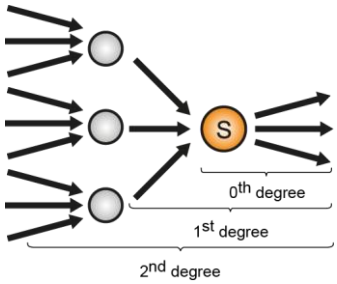


Figure 3.3 Schematic representation of three CoCCoA models with varying degree of network constraints included. 0^{th} degree CoCCoA model includes information from reactions consuming substrate *S* and fluxes through these reactions. 1^{st} degree CoCCoA model expands to additionally include information from reactions producing metabolite *S*. In the 2^{nd} degree CoCCoA, additional data from production reactions of direct producers of *S* is included in the model formulation to explain changes in abundance of *S*. Concentrations of neighbors of metabolite *S* (grey circles) are not needed in CoCCoA models.

0^{th} degree concentration change coupling analysis

We consider metabolite concentration changes relative to a reference condition – arbitrarily chosen from one of the two conditions pertaining to the experiment under investigation. Assuming that the

enzyme properties, represented by K_M, remain unchanged in the experiment, by using MM kinetics one obtains (Supplementary notes):

$$\frac{S}{S^*} = \left(\frac{V}{V^*}\right)\left(\frac{V^*_{\max} - V^*}{V_{\max} - V}\right) \quad (1)$$

Where * denotes the reference condition. The relative nature of this formulation allows circumventing the problem of the lack of availability of *in vivo* K_M values. Furthermore, by assuming that $V << V_{\max}$ & $V^* << V^*_{\max}$, and that the ratio $V^*_{\max} / V_{\max}$ can be approximated by the gene expression ratio, equation (1) simplifies to a log-linear relationship (equation 2, Supplementary notes). Both of these assumptions are critically examined below.

$$\ln\frac{S}{S^*} = \ln\frac{V}{V^*} - \ln\frac{T}{T^*} \quad (2)$$

The first assumption implies that the enzyme is not saturated – the opposite situation is not amenable for establishing the desired metabolite-gene expression (or enzyme availability in general) relationship, as the reaction velocity will then be only a weak function of metabolite concentration. Recent studies have shown that the *in vivo* concentration for several metabolites, esp. from central carbon metabolism, is close to the corresponding K_M values (Bennett et al, 2009). At these concentrations (Gerosa & Sauer, 2011) reaction rates V are close to half of the V_{max}. Although the assumption of $V<<V_{max}$ is not strictly applicable in this flux regime, numerical simulations show modest errors due to this approximation (around 20%, **Supplementary Figure 3.1**). Given the advantage that this approximation brings, namely elimination of need for knowing *in vivo* kinetic parameters, the cost of approximation error is rather low. This is an enabling assumption for linking transcriptome to metabolome, as currently no reliable estimates for *in vivo* kinetic parameters are available at the metabolic network scale.

The second major assumption used in the current analysis is the proportionality between mRNA fold-change and protein fold-change. Importantly, correlation between mRNA and protein abundances is not a necessity for our assumption, but only the correlation between the fold-changes of them across two conditions. This assumption is of particular relevance as the role of translation efficiency and post-translational modifications in regulating metabolic enzymes is becoming increasingly evident (Daran-Lapujade et al, 2007; Metallo & Vander Heiden, 2010; Oliveira & Sauer, 2012; Ptacek et al, 2005). We here note that the CoCCoA framework is readily applicable to proteomics data. Currently

there is lack of experimental data where both protein and metabolite abundances have been simultaneously measured in a network-wide manner. Nevertheless, we critically examined our hypothesis by analyzing published experimental data for *S. cerevisiae* where genome-wide mRNA and protein fold changes were simultaneously measured. Notably, correlations between mRNA and protein fold changes were found to be significantly stronger in the case of metabolic genes (Dataset from (Usaite et al, 2008a; Usaite et al, 2008b), $R^2 = 0.77$, $P = 0.04$; $R^2 = 0.66$, $P = 0.0365$, $R^2 = 0.76$, $P = 0.0036$; dataset from (Washburn et al, 2003), $R^2 = 0.4$, $P = 0.296$; dataset from (Ideker et al, 2001), $R^2 = 0.57$, $P = 0.0681$; dataset from (Griffin et al, 2002), $R^2 = 0.43$, $P = 0.0001$; *P*-values were estimated based on permutation test) (**Figure 3.4**A, **Supplementary Figure 3.2**). Interestingly, mRNA and protein changes have recently been demonstrated to be in good agreement even in mammalian systems (Schwanhausser et al, 2011).

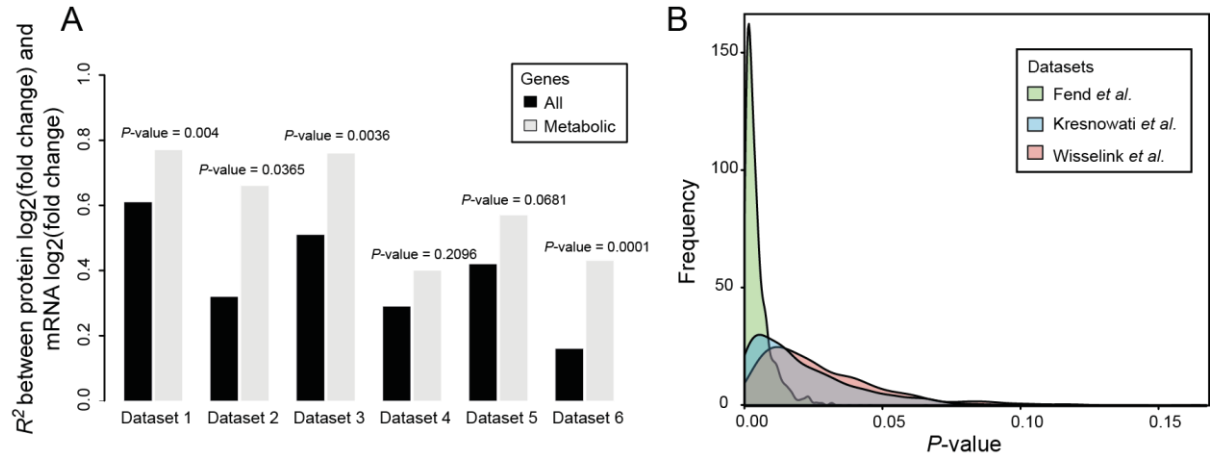


Figure 3.4 Correlation between protein abundance changes and the corresponding mRNA abundance changes is stronger for metabolic proteins. A) Black bars represent squared correlations including all proteins measured in each of the datasets. Grey bars are correlations only for metabolic proteins (as per genome –scale metabolic model by (Forster et al, 2003)). To assess the significance of obtained correlations, we calculated 10,000 different correlations between randomly chosen protein-transcript pairs (number of chosen pairs being equal to the number or metabolic protein measured in each dataset). *P*-value was estimated as a fraction of random correlations which were larger than those obtained for the metabolic proteins in the corresponding dataset. Each bar group represent different dataset (from left to right), 1,2,3 – (Usaite et al, 2008a; Usaite et al, 2008b), 4 – (Washburn et al, 2003), 5 – (Ideker et al, 2001), 6 – (Griffin et al, 2002). B) Distributions of *P*-values obtained for 1st degree CoCCoA analysis accounting for the variability in mRNA-protein relationship. For each case study, *P*-values were obtained from 1,000 different correlations between metabolite concentration changes and 1st degree CoCCoA score with correction factor for protein changes (see main text). Distributions are colored according to different datasets (Fendt et al, 2010; Kresnowati et al, 2006; Wisselink et al, 2010) used in the present work

The model represented by equation 2 is hereby termed as 0th degree coupling, meaning that the metabolite *S* is not coupled to any other metabolite and connected only to the enzyme using it as a substrate (**Figure 3.3**). We note that under the condition of flux homeostasis, *i.e.* no flux change between the two conditions, metabolite concentration ratio becomes dependent only on the transcript changes. The flux homeostasis model of the 0th degree coupling is equivalent to the

protein-metabolite analysis proposed by Sauer and co-workers (Fendt et al, 2010). In contrast to the analysis in (Fendt et al, 2010), significant correlation (r = -0.68, P = 0.021, n = 11) resulted from 0^{th} degree flux homeostasis analysis (**Figure 3.5**A), owing to the wider connectivity and hence the coverage of genes considered in the current analysis. In this case study, the observed physiology of the mutant showed 30% slower growth rate than the reference. Slower growth was concurrent with lower glucose uptake rate and implies large changes in intra-cellular fluxes spanning all major pathways. Indeed, we observed further improvement in the correlation by relaxing the assumption of flux homeostasis and thereby including the simulated flux data (**Figure 3.5**D) (Methods). For the other two datasets, however, the 0^{th} degree coupling model failed to explain the observed metabolic changes (**Figure 3.5**B, C, E, F). Thus, in case of lack of reliable flux estimates on the network-wide scale (which are currently difficult to obtain; and flux simulations based on limited physiological data do not give unique solutions due to network complexity), the 0^{th} degree model appears to be insufficient in terms of quantitatively connecting gene expression to metabolite levels.

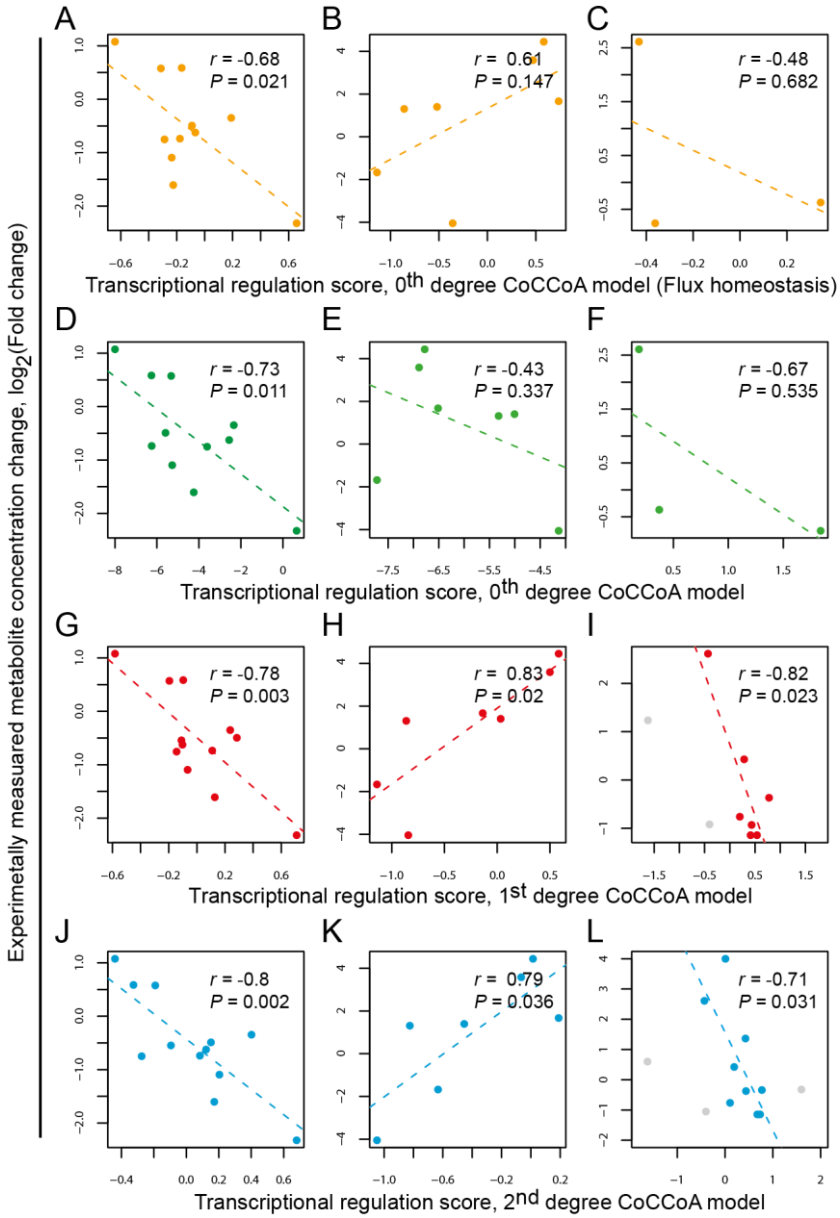


Figure 3.5 Metabolite concentration changes are explained by transcriptional regulation score based on concentration changes coupling analysis. A) 0^{th} degree CoCCoA with the assumption of flux homeostasis, data from (Fendt et al, 2010); B) data from (Kresnowati et al, 2006); C) data from (Wisselink et al, 2010). D) 0^{th} degree CoCCoA, data from (Fendt et al, 2010); E) data from (Kresnowati et al, 2006); F) data from (Wisselink et al, 2010). G) 1^{st} degree CoCCoA, data from (Fendt et al, 2010); H) data from (Kresnowati et al, 2006); I) data from (Wisselink et al, 2010). J) 2^{nd} degree CoCCoA, data from (Fendt et al, 2010); K) data from (Kresnowati et al, 2006); L) data from (Wisselink et al, 2010). Number of data points between coupling degree types can vary depending on number of connected genes (only significantly changed genes are considered, $\alpha = 0.05$, two tailed) to metabolite at certain degree. Metabolite data represent experimentally measured values (log_2-scale) as reported in the original study. Notice, for the 0^{th} degree model (D, E, F) values on abscissa axis are different order of magnitude comparing to all others, since for these CoCCoA scores include simulated flux change estimations. Data points and regression lines are colored according to the degree of concentration changes coupling: orange – 0^{th} degree (homeostasis assumption), green – 0^{th} degree, red – 1^{st} degree, cyan – 2^{nd} degree. Grey points (I, L) represents metabolites (glucose 6-phopshate, glucose 1-phosphate, alpha,alpha'-trehalose 6-phosphate) with which were not possible to include in the analysis due to altered flux directions when grown on two different carbon sources.

1st degree concentration change coupling analysis

In vivo, each metabolite pool is dependent on reaction consuming it as well as reactions producing it. At steady-state, sum of fluxes through the reactions that produce a particular metabolite, must be equal to the sum of the fluxes that use it as a substrate. For a metabolite with a single production reaction and a single consumption reaction, the steady-state assumption leads to equation (3) (Supplementary notes).

$$\ln\frac{S}{S^*} = \ln\frac{T^{prod}}{T^{prod*}} - \ln\frac{T^{cons}}{T^{cons*}} + \ln\frac{R}{R^*} \qquad (3)$$

T_{prod} and T_{cons} are gene expression levels corresponding to the enzymes producing and consuming *S*, respectively. *R* is concentration of pre-cursor metabolite of *S*. This relation implies metabolite concentration changes coupling between *R* and *S* and is defined as 1st degree coupling. Notably, the flux *V* is now eliminated and is accounted for by T_{prod} and *R*. Equation (3) brings a new network-perspective to the enzyme kinetics and thus provides a mechanistic basis for including information from the network connections and corresponding bio-molecular abundances The link between metabolite and transcriptional changes now includes components from both consumption and production reactions of that metabolite (**Figure 3.3**). When multiple reactions are consuming (or producing) the same substrate *S*, the consumption (or production) term can be approximated by the average transcript ratios of the all consumption (or production) reactions (proof is provided in Supplementary notes).

Using the above 1st degree concentration changes coupling model, we calculated correlation coefficients between metabolite concentration ratios and the 1st degree coupling score based on transcripts (first two terms on RHS of equation 3). Notably, we obtained significant correlations (P = 0.003, n = 12; P = 0.02, n = 7; P = 0.023, n = 7) in all three case studies (**Figure 3.5**G, H, I), as opposed to the lack of, or weak, correlations in the case of 0th degree coupling. It is important to note that the number of metabolites that can be assigned CoCCoA score can vary between the coupling degrees. Since only significantly (α = 0.05, two tailed) changed transcripts are considered, the number of transcripts that can be used for calculation typically increase as more distant reaction nodes in the network are included. For example, in case of arabinose case study (Wisselink *et al.*, 2010) (**Figure 3.5**C, F), only three metabolites have significant transcript changes corresponding to their consuming reactions and hence only these can be compared against the experimental data. In contrast, 1st degree CoCCoA scores can be calculated for 7 metabolites. We also note that the proposed analysis

excludes reactions for which the flux directions switch between the conditions being compared. Consequently, the corresponding neighboring metabolites cannot be assigned CoCCoA scores. For example, metabolites shown as grey circles (**Figure 3.5**C, F).

Network propagation of concentration control

In a similar way as going from the 0^{th} to the 1^{st} degree coupling, one can further extend the degree of network propagation of concentration control by replacing the concentration ratio in the RHS of Equation 3 with the 1^{st} degree CoCCoA relationship for the corresponding pre-cursor metabolites. In this fashion, we included the pre-cursor's production reactions to derive the second-degree coupling relationship (**Figure 3.3**) (Supplementary notes). Interestingly, the 2^{nd} degree correlations remained as strong as for the 1^{st} degree for the first two case studies. This result is notable since the use of 2^{nd} degree coupling involves expression data from the genes that are further away from the metabolites for which concentration is being predicted. Together, the results from the 1^{st} and the 2^{nd} degree coupling suggest that the majority of the control over concentration of a metabolite lies in its immediate neighbors in the network and the first-degree network connectivity appears to suffice in determining the fate of metabolite concentrations.

An important observation from our analysis is that the slopes of the correlations (**Figure 3.5**) are consistent with the corresponding overall flux change inferred from the physiological data. In case of (Fendt et al, 2010) data (**Figure 3.5**A, D, G, J), fluxes in the mutant strain were lower compared to the wild type due to reduced growth rate. The same observation applies to the glucose pulse case study (Kresnowati et al, 2006) (**Figure 3.5**H, K). Flux through the glycolytic reactions was increased following the pulse and this is reflected in the observed positive slope (**Figure 3.5**B). According to the proposed models, we would expect to have positive slopes independent of the flux changes. The negative slope likely results from several unknowns, from network connectivity and/or from the reaction mechanism itself. These factors include, among others, reaction directions, allosteric regulation and relationship between gene/protein abundance and enzyme activity (for example, resulting from post-translational modifications). The lack of network-wide data for *in vivo* enzyme activities and reaction directions currently limit our explanation of this interesting observation to empirical and qualitative nature. Nevertheless, CoCCoA provides a modeling framework to systematically integrate such data in the future and thereby towards building a more predictive model. In the following, we illustrate how uncertainty in correlations between mRNA and protein can be included into CoCCoA.

Metabolic genes in yeast show significantly stronger correlation between mRNA and protein fold-changes (**Figure 3.4**A, **Supplementary Figure 3.2**) in contrast to the other genes. Starting with the mRNA *vs.* protein fold change data for the yeast metabolic genes, we re-performed 1st degree CoCCoA whereas each transcript ratio was adjusted by using a randomly sampled correction factor. The space for this mRNA-protein fold change correction factor was estimated based on the slopes of linear regression lines between mRNA and protein fold changes across different datasets (**Supplementary Figure 3.2**, Supporting notes). We then examined whether the correlation between 1st degree CoCCoA scores and metabolite fold changes remained significant in 1,000 simulations using mRNA-Protein correction factors. For all of the three case studies (Fendt et al, 2010; Kresnowati et al, 2006; Wisselink et al, 2010), large fraction of these simulations (99%, 86%, 91%) remained significant (Fisher transformation, $\alpha = 0.05$) (**Figure 3.4**B), thereby indicating a high degree of robustness of CoCCoA towards different mRNA/Protein fold change ratios for different genes.

All three case studies examined above covered metabolites participating in the central carbon metabolism. In case of the certain enzymes from these and other pathways, for example amino acid metabolism, regulation by other mechanisms such as allosteric inhibition (Braus, 1991) may play a dominant role (Luttik et al; Metallo & Vander Heiden, 2010). Furthermore, in the case of enzymes saturated with their substrates, reaction rates (according to MM kinetics) will be decoupled from metabolite concentrations. When larger scale quantitative metabolomics data becomes available in the future, mechanisms in addition to those considered here, for example substrate inhibition, multi-substrate reaction mechanisms (Cleland, 1989) or allosteric regulation may need to be invoked in order to link metabolite concentrations with the enzyme abundance and hence to gene expression.

Discussion

One of the main observations from our results (**Figure 3.5**) is that the correlation between metabolites and transcripts becomes evident only when including network flow constraints. This suggests that the network connectivity component is important element that brings missing part to gene expression-metabolite puzzle. The use of genome-scale metabolic model allowed us to exploit the network connectivity element to a very high degree. For example, even for a sparsely connected metabolite such as D-Ribose 5-phosphate, the 2nd degree CoCCoA score accounts for transcriptional information from 17 genes (with significant change in gene expression), whereas 0th degree score accounts for only 2 transcripts (**Figure 3.6**A). Highly connected metabolic cofactors such as ATP and NADH also contributed to the observed correlations based on our results (**Figure 3.5**A, D, G, J,

Supplementary Figure 3.3). Given the importance of these hub metabolites in several human health and biotechnological problems (Brochado et al, 2010; Hahn-Hagerdal et al, 1996; Zelezniak et al, 2010)), CoCCoA represents a step towards understanding the transcriptional control of these pools. These results suggest a dominant contribution of the network flow constraints to the metabolite concentration control – in addition to the substrate-enzyme relationship given by the reaction mechanism. Expanding the network information beyond the first-degree neighbors appears to have a perturbation specific impact on the corresponding metabolite-gene expression relationship. As the degree of network connectivity in CoCCoA increases, larger numbers of new genes become part of the CoCCoA score for any given metabolite (**Figure 3.6**B). Importantly, inclusion of many new genes when moving from 1^{st} to 2^{nd} degree CoCCoA in general maintains the correlation (**Figure 3.5**). This observation suggests a strong co-regulation of genes that are linked through common substrates/products. Indeed, co-expression of metabolic genes on short network distances has been observed earlier (Kharchenko et al, 2005). CoCCoA thus provides a mechanistic explanation for the role of network topology in regulating global metabolite concentration changes. In particular, inclusion of production reactions (1^{st} degree CoCCoA) due to mass balance constraints makes a major contribution in explaining metabolite pool changes.

To what extent the discrepancy between mRNA and protein levels may explain the remaining variance in the metabolite-mRNA correlations predicted by CoCCoA? Single-experiment transcriptomics-proteomics datasets (Griffin et al, 2002; Ideker et al, 2001; Usaite et al, 2008a; Usaite et al, 2008b; Washburn et al, 2003) allowed us to examine the mRNA-protein correlations from the metabolism point of view. An interesting finding is that metabolic genes in yeast display significantly stronger correlation between mRNA and protein fold-changes as opposed to the genes involved in other cellular processes (**Figure 3.4**A, **Supplementary Figure 3.2**). It remains to be seen whether the enriched correlation between transcript and protein levels in metabolism extends to the enzyme activity. Increasing availability of genome-wide protein phosphorylation/acetylation data may aid in addressing this question. Any such information at protein abundance/activity level can be integrated in to the CoCCoA formulation in a straightforward manner – by substituting the corresponding transcript ratio terms by the protein abundance or activity measures. In light of the currently known variability of mRNA-protein ratios, CoCCoA remains valid when gene expression changes are corrected for the expected response at the level of proteins (**Figure 3.4**B).

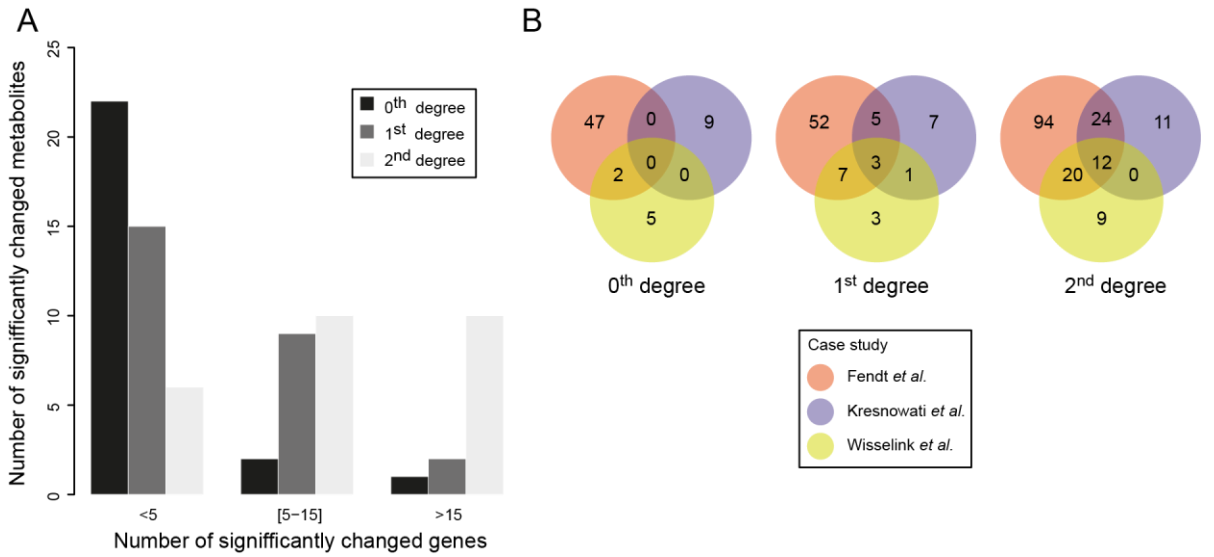


Figure 3.6 Gene coverage in CoCCoA case studies. A) Number of significantly changed genes increase with CoCCoA degree. B) Overlaps of significantly changed genes used for calculation of CoCCoA scores for all metabolites from different case studies. Number of genes that overlap across different case studies is relatively low, confirming the fundamentally different nature of the underlying perturbations. Case studies used in the present work: (Fendt et al, 2010; Kresnowati et al, 2006; Wisselink et al, 2010).

Formulation of CoCCoA in a relative manner, *i.e.* in the fold-change space, allows circumventing the problem of unknown *in vivo* kinetic parameters. A major advantage following this is that the CoCCoA models neither need any parameter fitting procedure nor need to be adjusted to any particular experiment. In contrast, kinetic models rely on parameter estimation from observed data, which makes model extrapolation across experiments difficult and dramatically increases the number of assumptions. Concentration change coupling analysis represents a step towards integrating kinetic and steady-state modeling approaches at the network-scale.

CoCCoA models do not require time-course experiments and can be applied to studies designed as comparison between two factors, *e.g.* wild-type *vs.* mutant cells. Indeed, the applicability of CoCCoA was found to be quite general in terms of the experimental design underlying the data. Our results demonstrate the applicability for three fundamentally different and biologically relevant perturbations – mutant/wild-type, environmental perturbation and adaptive evolution. These three case studies also comprise of two biologically different growth conditions, batch (Fendt et al, 2010) and chemostat (Kresnowati et al, 2006; Wisselink et al, 2010). The difference in the nature of these perturbations is reflected in the fact that despite the concentration changes within the same metabolic pathways are detected in these studies; gene expression changes have only small overlap (**Figure 3.6**B). This attests the generality of the CoCCoA model as well as empirically shows its robustness for application to different biological perturbations.

CoCCoA models explain more than 60% of the variation in metabolite changes based on changes in expression of the genes coding for producer and consumer enzymes of these metabolites. This result strongly suggests that the proposed (log-) linear relationships between metabolite and gene expression changes are biologically meaningful. Quantitative predictions for the slopes of the corresponding regression lines will be a major challenge for the future modeling work. In the current work, we observe that the sign of the CoCCoA-metabolite relationship matches with the up/down regulation of fluxes from that of the reference state.

Application of CoCCoA is not possible in the case of perturbations that are likely to drastically affect properties of several enzymes. Furthermore, metabolite-enzyme pairs where metabolite concentrations are likely to be above their K_M values (Bennett et al, 2009; Gerosa & Sauer, 2011) the relationship between metabolite and enzymes will be largely independent of changes in enzyme abundances. In higher eukaryotes such as plants or mammalian cells (Urbanczyk-Wochniak et al, 2003), potential influence of compartmentalization, feedback regulation, among others, will have important effects and new methods need to be developed to address these. For example, genome-wide metabolite profiling in plants and their associations with genomic and phenotypic traits can provide a valuable starting points towards this (Schauer et al, 2006).

In conclusion, concentration change coupling analysis yielded so far best-known correlations between gene expression and metabolite levels. CoCCoA provides a framework towards bridging the gap between functional genotype and the observed metabolic phenotype. In the yeast central carbon metabolism, this bridge appears to be supported on the transcriptional regulation operating around the metabolites.

Methods

Datasets

The first dataset (Fendt et al, 2010) consists of a comparison of a reference yeast strain with GCR2 null mutant. GCR2 is a transcription factor responsible for activation of glycolytic genes (Uemura & Jigami, 1992). In the second case study (Kresnowati et al, 2006), *S. cerevisiae* was grown in carbon-limited chemostat cultures and was subjected to a pulse change in glucose concentration. In the third dataset (Wisselink et al, 2010), an evolutionarily engineered yeast strain was grown on glucose or arabinose as a sole carbon source. Summary of growth conditions and descriptions of datasets from all case studies is provided in **Supplementary Table 3.3**. Metabolite data was used as reported in the

original studies; significance cut-off α = 10% was chosen to control for type 1 error of null hypothesis that mean metabolite concentrations between two conditions are equal.

Metabolic network and flux variability analysis

Genome-scale reconstruction of *Saccharomyces cerevisiae* metabolic network by (Forster et al, 2003) was used to link metabolites with the enzyme-coding genes. A total of 708 structural open reading frames (ORFs) are accounted for in this network, representing 1035 metabolic reactions (in total 1175 reactions and 584 metabolites). For each of the case studies, reaction directions were estimated by using flux variability analysis (Mahadevan & Schilling, 2003). The steady-state model was constrained with the corresponding physiological data as reported for the corresponding case studies (**Supplementary Table 3.1**, **Supplementary Table 3.2**, **Supplementary Table 3.3**, **Supplementary Table 3.4**, **Supplementary Table 3.5**). Resulting linear programming problems were solved using *glpk* (http://www.gnu.org/software/glpk/) solver accessed through a C library. Flux fold-change data required for the 0^{th} degree concentration changes coupling scores were estimated as ratios of average of the minimum and maximum feasible flux before and after perturbation.

Transcription data analysis

Preprocessing of the Affymetrix CELL files was carried out by using the statistical software environment R/Bioconductor (http://www.bioconductor.org). Resulting probe intensities were corrected for background by using robust multi-array average method (Irizarry et al, 2003 2003) (RMA) using only perfect-match (PM) probes. Normalization was performed by using the quantiles algorithm. Gene expression intensity values were calculated from the PM probes using median polish summarization method (Irizarry et al, 2003 2003). Significance of the differential expression was calculated by using the empirical Bayes test implemented in *limma* package (Smyth, 2004).

Statistical analysis

Pearson correlation coefficients between $\log_2$ metabolite fold changes and concentration changes coupling scores were calculated with statistical software R (http://www.r-project.org) by using linear model fitting function lm. Metabolite changes were used as dependent variables and CoCCoA scores as independent. *P*-values for null hypothesis of no correlation (regression slope = 0) were estimated by applying anova function to linear model (*lm*) object. In addition, we performed a permutation test for correlations between metabolite fold changes ($\log_2$) and CoCCoA scores. The originally paired data was randomly permuted without replacement for 10,000 times. For each permutation, a correlation coefficient was calculated and the *P*-value was estimated as a fraction of squared

correlation coefficients which were larger than in the original paired data. The results were similar to those estimated by *anova* function.

Acknowledgements

We thank T. Çakir, J. Nielsen, A.R. Brochado and Z. Soons for constructive feedback on the manuscript. We thank R. Olivares-Hernandez for help with compilation of the proteomics data.

Contributions

K.R.P. and A.Z. designed the study, developed the approach and formulated the model. A.Z. implemented the algorithm and performed data analysis. K.R.P. and A.Z. wrote the manuscript.

Conflicts of interest

The authors declare that they have no conflict of interest.

Supplementary information

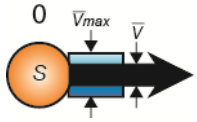
Contents

List of Supplementary Figures

List of Supplementary Tables

Supplementary notes

0th degree concentration change coupling



Michaelis-Menten kinetics equation is given by:

$$V = \frac{V_{\max} S}{K_M + S}$$

Rearranging for *S*:

$$S = \frac{V K_M}{V_{\max} - V} \qquad (1)$$

For reference condition:

$$S^* = \frac{V^* K_M}{V^*_{\max} - V^*}$$

Dividing Eq.1 by reference condition gives:

$$\frac{S}{S^*} = \left(\frac{V}{V^*}\right)\left(\frac{V^*_{\max} - V^*}{V_{\max} - V}\right) \qquad (2)$$

Assuming that $V << V_{max}$ & $V^* << V_{max}^*$ (see main text for assumption discussion) one gets:

$$\frac{S}{S^*} = \frac{V}{V^*}\frac{V^*_{\max}}{V_{\max}} \qquad (3)$$

Transforming to log-space Eq. 3 becomes:

$$\ln\frac{S}{S^*} = \ln\frac{V}{V^*} + \ln\frac{V^*_{\max}}{V_{\max}}$$

Since $V_{max} = k_2[E]$, where k_2 and $[E]$ are substrate to product conversion rate and concentration of enzyme of active enzyme, respectively. Assuming that $E \propto T$, there T is transcript abundance, and that k_2 is not changing between two conditions, one gets:

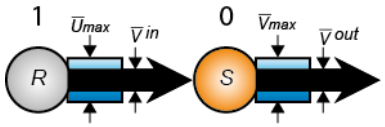
$$\ln\frac{S}{S^*} = \ln\frac{V}{V^*} + \ln\frac{T^*}{T}$$

or

$$\ln\frac{S}{S^*} = -\ln\frac{T}{T^*} + \ln\frac{V}{V^*} \qquad (4)$$

Eq. 4 is defined as 0[th] degree concentration change coupling.

1st degree concentration change coupling§



At steady-state V_{in} =$V_{out,}$ consequently:

$$\frac{V_{in}}{V_{in}^*} = \frac{V_{out}}{V_{out}^*}$$

Substituting flux ratio in above equation by metabolite and transcript ratios as given by Eq.4, we obtain:

$$-\ln\frac{R}{R^*} - \ln\frac{T_{cons}^R}{T_{cons}^{R\ *}} = -\ln\frac{S}{S^*} - \ln\frac{T_{cons}}{T_{cons}^*},$$

Since the consuming reaction of R is the same as the production reaction of *S*, $T_{cons}^R = T_{prod}$.

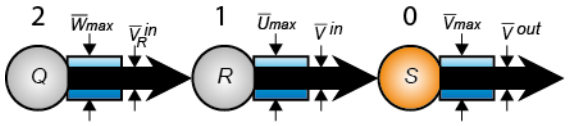
Rearranging for *S* thereby gives:

$$\ln\frac{S}{S^*} = -\ln\frac{T^{cons}}{T^{cons*}} + \ln\frac{T^{prod}}{T^{prod*}} + \ln\frac{R}{R^*} \qquad (5)$$

Eq. 5 is defined as 1[st] degree concentration change coupling.

§In presented scheme, U_{max} is equivalent of V_{prod}^{max}

2nd degree concentration change coupling§



Applying first degree CoCCoA (Eq.5) to metabolite *R*, gives:

$$\ln\frac{R}{R^*} = -\ln\frac{T_R^{cons}}{T_R^{cons*}} + \ln\frac{T_R^{prod}}{T_R^{prod*}} + \ln\frac{Q}{Q^*} \qquad (6)$$

Since $\ln\frac{T_R^{cons}}{T_R^{cons*}} = \ln\frac{T^{prod}}{T^{prod*}}$, therefore by substituting *R* term from Eq.5 into Eq.6 gives:

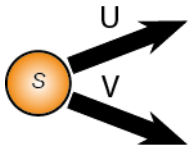
$$\ln\frac{S}{S^*} = -\ln\frac{T^{cons}}{T^{cons*}} + \ln\frac{T_R^{prod}}{T_R^{prod*}} + \ln\frac{Q}{Q^*} \quad (7)$$

Eq. 7 is defined as 2nd degree concentration change coupling.

§In presented scheme, U_{max} is equivalent of V_{prod}^{max} and W_{max} is related to the production of metabolite R

Multiple reactions connected to S

0th degree concentration change coupling



Two reaction using the same substrate *S*. Only for consumption reactions 0th degree concentration change coupling. *U* and *V* represent fluxes through the two reactions.

Let's consider two reaction carrying flux *U* and *V* use the same substrate *S*. For each reaction, 0th degree concentration change coupling (Eq.4) can be applied independently

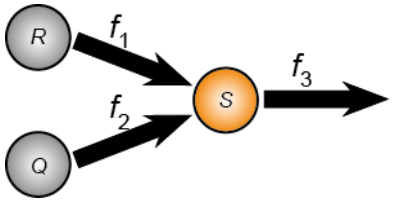
$$\begin{cases} \ln\frac{S}{S^*} = -\ln\frac{T_U}{T_U^*} + \ln\frac{U}{U^*} \\ \ln\frac{S}{S^*} = -\ln\frac{T_V}{T_V^*} + \ln\frac{V}{V^*} \end{cases}$$

Above system of two equations can be summed and rearranged as:

$$\ln\frac{S}{S^*}=\frac{1}{2}\left(-\ln\frac{T_U}{T_U^*}+\ln\frac{U}{U^*}-\ln\frac{T_V}{T_V^*}+\ln\frac{V}{V^*}\right) \qquad (8)$$

For more than two reactions, similar analysis will imply averaging of fold changes of the corresponding transcripts; see also **Figure 3.1**B in the main text.

1st degree concentration change coupling



Multiple reactions producing same metabolite, 1st degree concentration change coupling

At steady-state:

$$f_1+f_2=f_3$$

Comparing to the reference:

$$\frac{f_1+f_2}{f_1^*+f_2^*}=\frac{f_3}{f_3^*}$$

$$\ln\left(\frac{f_1+f_2}{f_1^*+f_2^*}\right)=\ln\frac{f_3}{f_3^*}$$

Let's define $f_1=\alpha f_2$ (or $f_2=\beta f_1$) correspondingly for reference condition, then:

$$\ln\left(\frac{f_1(1+\alpha)}{f_1^*(1+\alpha^*)}\right)=\ln\frac{f_3}{f_3^*}$$

It was shown that flux ratio, for the most of perturbations, do not change (Haverkorn van Rijsewijk et al, 2011), thus we can consider $\alpha=\alpha^*$ and $\beta=\beta^*$, *i.e.* the split ratio of fluxes between conditions is not changing, then:

$$\begin{cases}\ln\dfrac{f_1}{f_1^*}=\ln\dfrac{f_3}{f_3^*}\\ \ln\dfrac{f_2}{f_2^*}=\ln\dfrac{f_3}{f_3^*}\end{cases}$$

$$\ln\frac{f_3}{f_3^*}=\frac{1}{2}\left(\ln\frac{f_2}{f_2^*}+\ln\frac{f_1}{f_1^*}\right) \qquad (9)$$

Rearrangement of Eq.9 using Eq.4 gives:

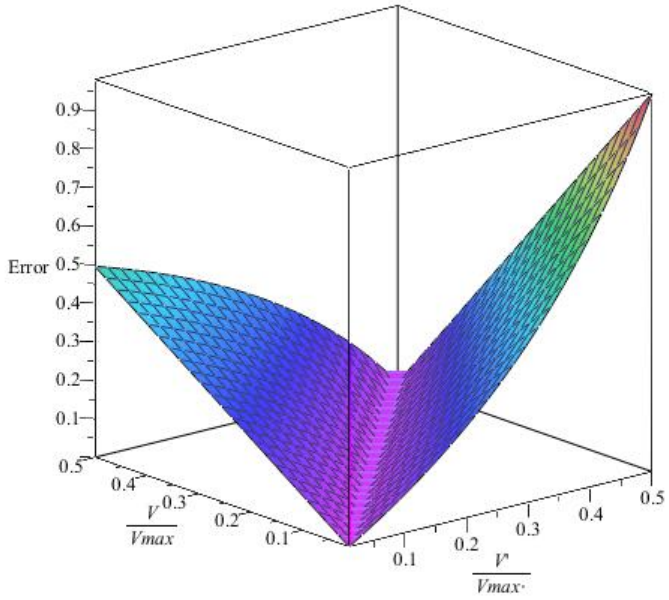
$$\ln\frac{S}{S^*}=\frac{1}{2}\left(\ln\frac{T_{f1}}{T_{f1}^*}+\ln\frac{R}{R^*}+\ln\frac{T_{f2}}{T_{f2}^*}+\ln\frac{Q}{Q^*}\right)-\ln\frac{T_{f3}}{T_{f3}^*} \qquad (10)$$

Equation 10 was used as a basis for calculation of 1st degree concentration change couplings; see also **Figure 3.4**B in the main text.

1st degree concentration change coupling with protein-mRNA correlation correction factor

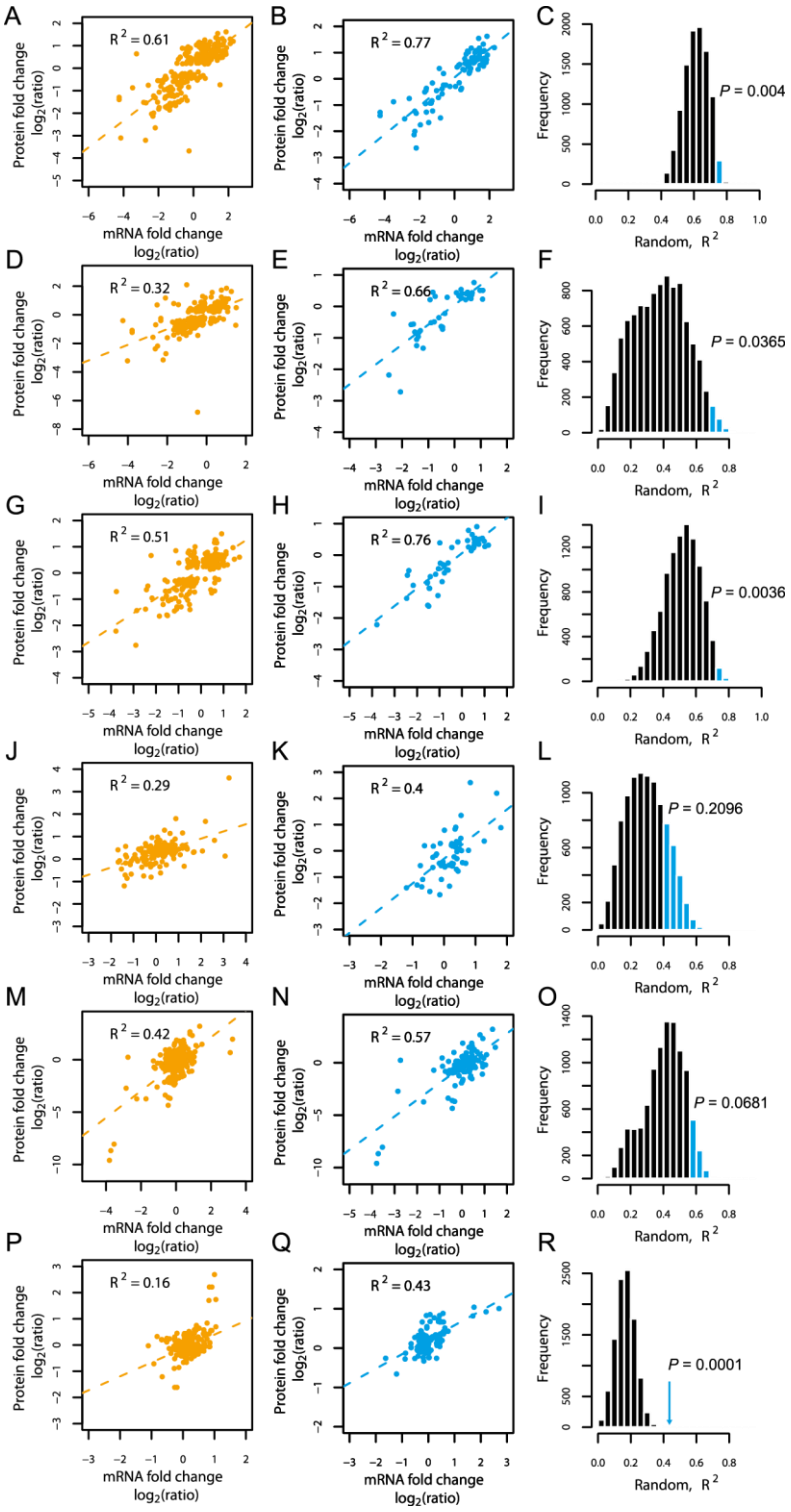
$$\ln\frac{S}{S^*}=-\beta_{cons}\ln\frac{T^{cons}}{T^{cons*}}+\beta_{prod}\ln\frac{T^{prod}}{T^{prod*}}+\ln\frac{R}{R^*}$$

Each transcript change term $\ln\frac{T}{T^*}$ was multiplied by correction factor β, which was estimated as a regression slope coefficient randomly sampled from normal distribution with mean and variance determined from least squares regression lines slopes of metabolic mRNA-protein fold changes correlations (**Supplementary Figure 3.2**). Data from studies (Griffin et al, 2002; Usaite et al, 2008a; Usaite et al, 2008b) with *P*-value < 0.05 (**Supplementary Figure 3.2**) was used for making distribution of slopes.

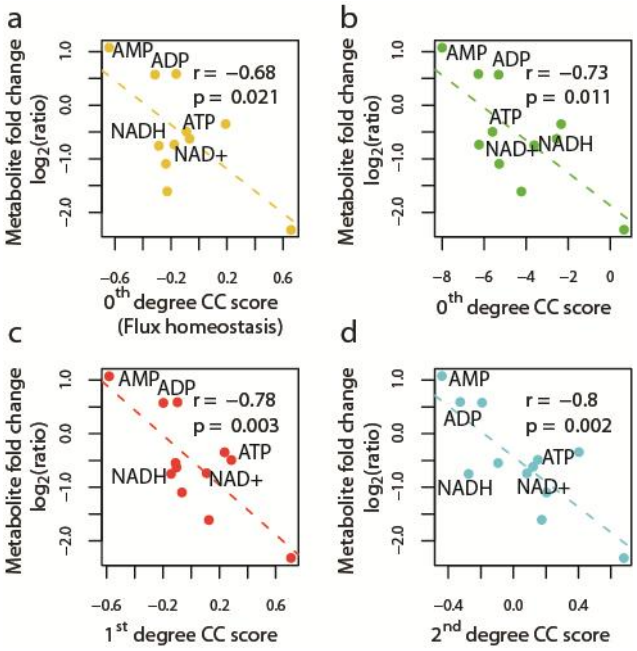


Supplementary Figure 3.1 Metabolite concentration change error as a function V/V_{max}(respectively V^*/V_{max}^*). In a standard Michaelis-Menten kinetics mechanism, K_M is a concentration there enzyme is half-saturated and reaction rate $V = 0.5\ V_{max}$ When substrate concentration reaches its K_M value (at reaction rates $0.25V_{max}$-$0.5V_{max}$), in order to explain metabolite concentration changes one needs to consider V (respectively V^*). In low flux ranges $0.1V_{max}$-$0.25V_{max}$ (respectively $0.1V_{max}^*$-$0.25V_{max}^*$) which is of order magnitude of K_M, metabolite concentration change can be explained under assumption ($V<<V_{max}$) with reasonable error. Error is defined by equation 11.

$$E = \sqrt{\left(1 - \frac{\frac{V}{V^*}\frac{{V_{max}}^*}{V^*}}{\frac{V}{V^*}\frac{{V_{max}}^*}{V^*}\frac{(1-\frac{V^*}{{V_{max}}^*})}{(1-\frac{V}{V_{max}})}}\right)^2} = \sqrt{\left(1 - \frac{1 - V/V_{max}}{1 - V^*/{V_{max}}^*}\right)^2} \qquad (11)$$



Supplementary Figure 3.2 Correlation between protein abundance changes and the corresponding mRNA abundance changes is stronger for metabolic proteins. A, D, G, J, M, P) Correlation including all proteins measured in each of the datasets. B, E, H, K, N, Q) Correlation only for the metabolic proteins (as per genome-scale metabolic model by (Forster et al, 2003)) measured in each of the datasets. C, F, I, L, O, R) Histogram of correlation coefficients obtained for 10,000 different correlations between randomly chosen protein-transcript pairs (number of chosen pairs being equal to the number of metabolic proteins measured in each dataset). Blue area denotes number of times when random correlations were higher than those obtained for the metabolic proteins in the corresponding dataset. Each row of plots represents a different dataset (from top to bottom), 1, 2, 3 –(Usaite et al, 2008a; Usaite et al, 2008b) ; 4 –(Washburn et al, 2003); 5 –(Ideker et al, 2001); 6 –(Griffin et al, 2002)



Supplementary Figure 3.3 Concentration change coupling for hub metabolites. Correlations between metabolite changes and concentration change coupling (CoCCoA) scores (derived from gene expression data). Data from (Fendt et al, 2010). The plots are the same as in **Figure 3.5** (main text), except that the highly connected metabolites are marked. a) 0^{th} degree CoCCoA with the assumption of flux homeostasis; b) 0^{th} degree CoCCoA; c) 1^{st} degree CoCCoA; d) 2^{nd} degree CoCCoA. Data points and regression lines are colored according to the degree of concentration change coupling: orange – 0^{th} degree (homeostasis assumption), green – 0^{th} degree, red – 1^{st} degree, cyan – 2^{nd} degree.

Supplementary Table 3.1 Summary of growth condition from all case studies used in analysis. Results from case studies 1, 2 and 3 are described in main text.

Study #	Reference	Growth conditions	Short description	Comparison	Remarks
1	(Fendt et al, 2010)	Batch (shake flask), aerobic, glucose minimal medium	Compares Gcr2p null mutant with reference yeast strain. Gcr2p responsible for activation of glycolysis. Measurements are taken at exponential growth phase.	Δgcr2 mutant *vs.* wild-type yeast	Growth rate of mutant was 30% slower than wild-type FY4.
2	(Kresnowati et al, 2006)	Chemostat, aerobic, glucose limited	Growing glucose-limited chemostat culture is subjected to glucose pulse. Data is collected at different time points before and after pulse.	300s after glucose pulse *vs.* time before pulse 0s	Time point of 300s was chosen to account for time needed for transcription to affect metabolite levels.
3	(Wisselink et al, 2010)	Chemostat, anaerobic, arabinose and glucose limited	Evolutionary adapted strain which is able to grow on arabinose was grown on arabinose or glucose.	Growth on arabinose of adapted *vs.* growth on glucose of adapted strain	For 0^{th} degree concentration change coupling only 3 points could be potentially used for correlation analysis (no significantly changed transcripts in consumption reactions)

Supplementary Table 3.2 Physiological data from case study 1 (Fendt et al, 2010). Comma-separated values denote lower and upper bounds used for constraining the corresponding fluxes.

Reaction (mmol/g/h)	***Δgcr2***	**WT**
Glucose uptake	10, 11	16.15, 17.85
Ethanol secretion rate	14, 15	20, 22
Glycerol secretion rate	1.66, 2.07	2, 2.2
Acetate secretion rate	0.5,0.7	0.75, 1.51
Pyruvate secretion rate	0.04, 0.06	0.06, 0.09
Growth rate	0.2,0.26	0.3, 0.33

Supplementary Table 3.3 Physiological data from case study 2 (Kresnowati et al, 2006). Parameters were estimated for 300s time point by using the plots provided in the original publication. Comma-separated values denote lower and upper bounds used for constraining the corresponding fluxes constraints.

Reaction (mmol/g/h)	**300s**	**0s**
Glucose uptake	3.6, 3.8	0.45, 0.6
Oxygen uptake	4.2, 4.4	1.55, 1.8
Ethanol secretion rate	2.75, 2.9	-
Acetate secretion rate	0.5, 0.6	-
Glycerol secretion rate	0.14, 0.16	-
Growth rate	1.66, 2.07	0.04, 0.048

Supplementary Table 3.4 Physiological data from case study 3 (Wisselink et al, 2010). Comma-separated values denote lower and upper bounds used for constraining the corresponding fluxes constraints.

Reaction (mmol[CmolDW]$^{-1}$ h^{-1})	**Arabinose**	**Glucose**
"Arabinose"*/Glucose uptake	61, 62	59, 63
Ethanol secretion rate	91.1, 104.5	96, 108
CO_2 secretion rate	97, 110	100, 112.8
Succinate secretion rate	0.29, 0.37	0.2, 0.5
Glycerol secretion rate	6.85, 7.15	6.9, 7.9
Acetate secretion rate	0.45, 0.81	0.376, 0.42
Pyruvate secretion rate	0.046, 0.054	0.066, 0.074
Growth rate	1.6, 1.66	1.45, 1.55

*Growth on glucose was used to estimate reaction directions. Note: CoCCoA cannot be used if many reaction directions are changing between the two conditions tested. In this case, however, according to (Wisselink et al, 2010), the difference in reaction directions are confined only for few fluxes. This enabled us to perform CoCCoA.

Supplementary Table 3.5 Changes in reaction directions used in case study 1 and 2. Glycolytic flux directions were fixed according to (Fendt et al, 2010). For other reactions, directions based on other literature indications were used.

Reaction name in the model	Reaction	Remarks
Glycolysis		
PGI1_1	alpha-D-Glucose 6-phosphate -> beta-D-Fructose 6-phosphate	
PGI1_2	alpha-D-Glucose 6-phosphate -> beta-D-Glucose 6-phosphate	
PGI1_3	beta-D-Glucose 6-phosphate -> beta-D-Fructose 6-phosphate	
FBA1	beta-D-Fructose 1,6-bisphosphate -> D-Glyceraldehyde 3-phosphate + Glycerone phosphate	
TDH1|TDH2|TDH3	D-Glyceraldehyde 3-phosphate + NAD+ + Orthophosphate -> 3-Phospho-D-glyceroyl phosphate + NADH	
PGK1	3-Phospho-D-glyceroyl phosphate + ADP -> 3-Phospho-D-glycerate + ATP	
GPM1_1	3-Phospho-D-glyceroyl phosphate -> 2,3-Bisphospho-D-glycerate	
GPM1_2|GPM2|GPM3	3-Phospho-D-glycerate -> 2-Phospho-D-glycerate	
ENO1|ENO2|ERR1_1|ERR1_2|ERR2	2-Phospho-D-glycerate -> Phosphoenolpyruvate	
ACO1|YJL200C	CitrateM -> IsocitrateM	
LSC1	ATPM + CoAM + ItaconateM -> ADPM + Itaconyl-CoAM + OrthophosphateM	
LSC2	ATPM + CoAM + SuccinateM -> ADPM + OrthophosphateM + Succinyl-CoAM	
FUM1_1	FumarateM -> MalateM	
FUM1_2	Fumarate -> Malate	
Amino acid metabolism		
SHM1	L-SerineM + TetrahydrofolateM -> 5,10-MethylenetetrahydrofolateM + GlycineM	
SHM2	L-Serine + Tetrahydrofolate -> 5,10-Methylenetetrahydrofolate + Glycine	
GDH2	L-Glutamate + NAD+ -> 2-Oxoglutarate + NADH + NH3	Reaction removed
OAC1	Oxaloacetate -> H+M + OxaloacetateM	
DIC1_1	Malate + SuccinateM -> MalateM + Succinate	
YDR111C	L-Glutamate + Pyruvate -> 2-Oxoglutarate + L-Alanine	
YFL030W	Glycine + Pyruvate -> Glyoxylate + L-Alanine	
CTP1_1	Citrate + MalateM -> CitrateM + Malate	

Chapter 4
Network architecture imparts plasticity to transcriptional regulation in the yeast metabolic network

Introduction

Studies of structural organization of metabolic network have provided a first at glance of properties and organization of metabolic networks (Albert, 2005; Csete & Doyle, 2004). Although information about topological characteristics is very useful, it provides only a static description of biological networks (Kharchenko et al, 2005). In reality, metabolic networks are subject to tight regulation at a number of levels including, transcriptional regulation. Previous studies have shown that functionally connected genes are often co-regulated (Gavin et al, 2002; Hartwell et al; Lee et al, 2002). Earlier observations suggested that co-regulated genes are often arranged in a linear order (Ihmels et al, 2004). Moreover, it was demonstrated that in metabolic networks, closely connected genes display significant co-regulation (Kharchenko et al, 2005). In our previous study (**Chapter 3**) we show that changes of metabolite concentrations to some extend are explained by difference of average changes in expression of genes neighboring it. By applying same principles (see **Chapter 3**), here, we demonstrate several novel aspects of transcriptional regulation in metabolic network of *Saccharomyces cerevisiae*.

Using gene expression data from multiple environmental/genetic perturbation experiments (Materials and Methods) we show that expression of genes centered around a metabolite can be regulated in three different ways. While some metabolites show a strong co-expression in only a few of its producing enzymes and consuming enzymes, their collective action tends to be coordinated. We suggest a new emergent type of regulation which accounts for all incoming and outgoing genes for a given metabolite. We show that emergent co-regulation is dominant among many metabolites in the metabolic network of *S. cerevisiae*.

Results and discussion

Previous studies have investigated the co-regulation of reaction around a metabolite by computing the correlation coefficients of the changes in gene expressions for corresponding enzymes across many experiments (Ihmels et al, 2004; Kharchenko et al, 2005). Earlier data suggested that metabolic genes in individual pathways were co-expressed, and as such, at divergent pathway branch points (for a given metabolite incoming reaction and two outgoing reactions), the expression of the incoming reaction was seen in the majority of cases to be correlated with only one of the two outgoing reactions (Ihmels et al, 2004). The past studies did not consider more than two incoming-outgoing reaction pairs at once. However, in our previous work (**Chapter 3**), we showed that changes of metabolite concentration are correlated with collective changes of genes neighboring it. Therefore, we hypothesize that metabolite concentration are dependent on overall action of enzymes neighboring it and expression of these enzymes are coordinated. In other words, to maintain homeostasis of the system and control metabolite concentration levels, changes of the metabolite's producing (input) genes must be co-regulated with consuming (output) genes.

To systematically examine the transcriptional regulation of the metabolic network of *S. cerevisiae,* we assembled a large gene expression dataset (Materials and Methods) spanning a variety of environmental and genetic perturbations experiments. In the metabolic network of *S. cerevisiae* (Forster et al, 2003) 34% of reactions are reversible; to ensure which metabolites are being produced/consumed, we therefore applied a criteria that the carbon source used in selected experiments had to be of glycolytic nature (Materials and Methods).

For all metabolic genes a fold change value can be calculated for each comparison between groups of samples in each experiment (Materials and Methods). By comparing the fold changes of each comparison for all input-output genes, a correlation coefficient can be calculated to measure the degree of co-regulation for the pair of genes. Alternately, the fold changes of all incoming reactions can be averaged for each comparison, and compared against the average fold change of all outgoing reactions for each comparison, to determine a single correlation coefficient for each metabolite, representing the collective co-regulation of all incoming and all outgoing reactions (**Figure 4.1**A). By computing correlation coefficients for all gene pairs, we found that coefficients of all in-out pairs are slightly higher than a control of random gene pairs (*P*-value < 10^{-7}, Wilcoxon rank sum test) (**Figure 4.1**B). In contrast, the correlation coefficients of the averages of production reactions *vs.* consumption reactions overall are noticeably higher than a control in which gene identities are shuffled (*P*-value < 10^{-7}, Wilcoxon rank sum test) (**Figure 4.1**C).

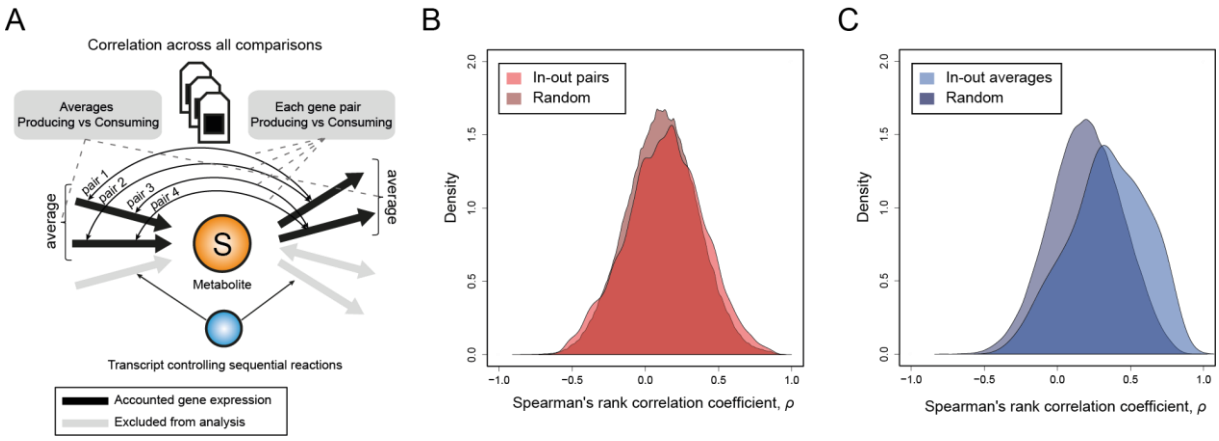


Figure 4.1 Correlation between metabolite neighbor genes. A) Metabolite can be a product of reactions (input reactions of metabolite S) and can participate in reactions as a substrate (output reactions of metabolite S). By comparing the fold changes in each comparison for all input-output genes pairs (those which code for the corresponding enzymes), a correlation coefficient is calculated to measure the degree of co-regulation for the pair of genes. Similarly, average of fold changes of inputs genes is correlated against average of all output genes across all comparison. To avoid autocorrelation transcripts coding enzymes performing sequential reactions were removed from analysis. After performing flux variability analysis, remained reversible reactions were removed from calculations (Materials and Methods). B) Distribution of the correlation coefficients between input *vs.* output gene pairs. For comparison, a random control was constructed by shuffling the identities of each gene 1000 times while preserving the connectivity of each metabolite. Importantly, only genes coding for enzymes were used in the random control. C) Distribution of correlation coefficients of average input genes *vs.* average output. Random control was computed as in B panel.

Subsequently, for each metabolite we compared correlation patterns of all producing-consuming gene pairs with correlations of averages expression changes of production-consumption genes (**Supplementary Figure 4.1**). A number of metabolites (*e.g.* NADPH, alpha-D-Glucose, Acetaldehyde) showed a high correlation for certain in-out pairs. In contrast, co-expression of consuming and producing genes was very low for these metabolites, suggesting a linear pathway-oriented regulation (**Supplementary Figure 4.1**). Another group of metabolites (*e.g.* Glycogen, Ubiqunol mitochnodrial, S−Adenosyl−L−homocysteine) showed a uniform regulation irrespective of its pathway, which was displayed by generally high correlation for in-out pairs, with a corresponding high co-regulation seen in producing-consuming averages. Strikingly, correlations of averages of producing *vs.* consuming genes were higher than in any individual producing-consuming gene-pairs for a considerable amount of metabolites (~15 %) showing an emergent type of regulation. **Figure 4.2** illustrates three potential regulatory schemes.

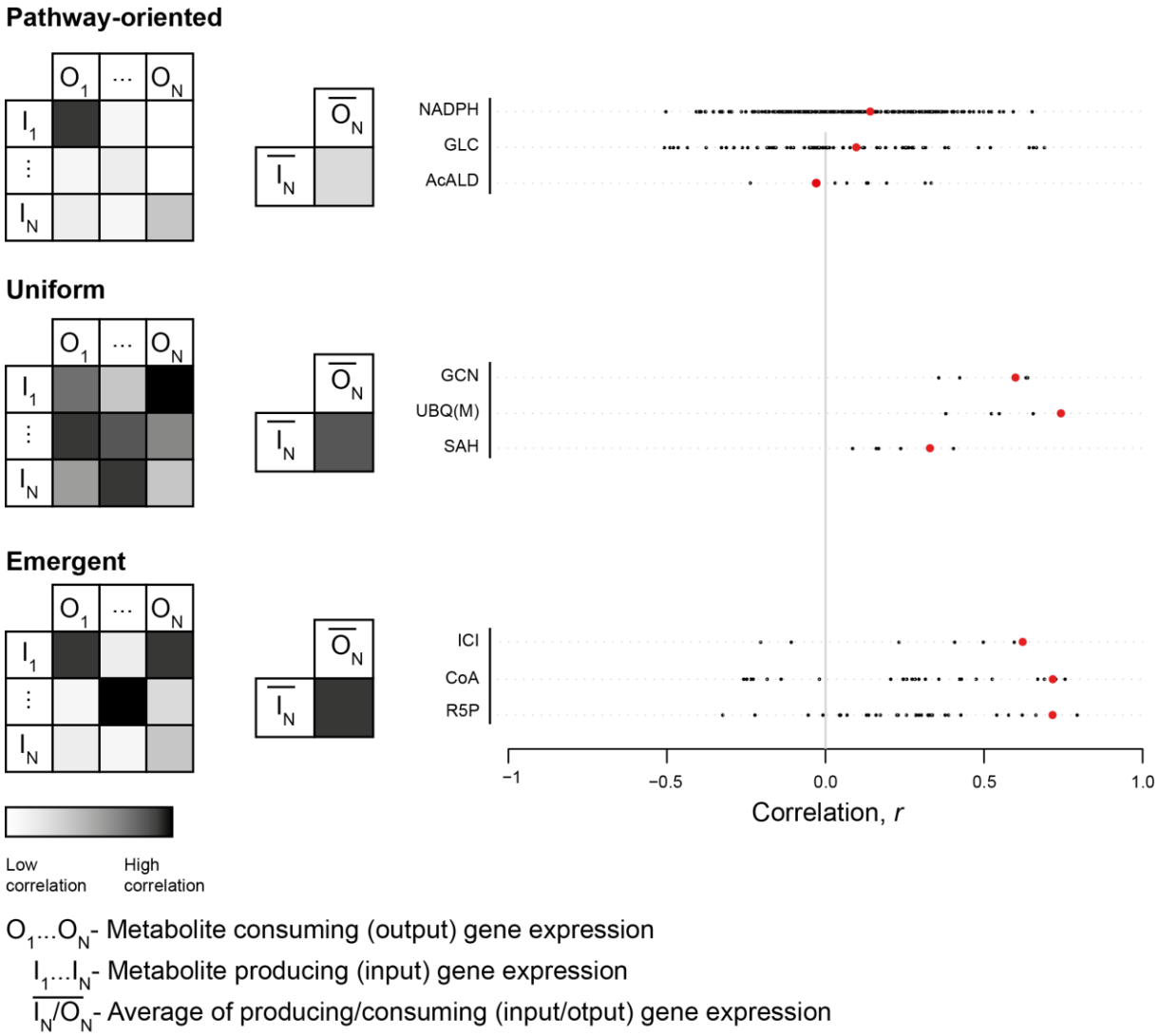


Figure 4.2 Three potential regulatory schemes and their signature co-regulation patterns. Linear pathway-oriented regulation would have a sparse highly-correlated in-out pairs, and low average-in-average-out correlation. Uniform metabolite-oriented regulation would show generally high correlations for in-out pairs, with a correspondingly high average-in-average-out correlation. Emergent regulation would have a variety of low-to-moderate correlations of in-out pairs, with an average-in-average-out correlation above the average correlation of pairs or above the maximum correlation of pairs. NADPH – Nicotinamide adenine dinucleotide phosphate; GLC – alpha-D-Glucose; AcALD – acetaldehyde; GCN – glycogen; UBQ(M) – Ubiquitinol mitochondrial; SAH – S−Adenosyl-L-homocysteine; ICI – isocitrate; CoA – Coenzyme A; R5P – ribose 5-phosphate.

The proposed uniform regulation of incoming reactions and outgoing reactions fits with our finding of strong correlations between average incoming fold changes with average outgoing fold changes (**Figure 4.1**C). This perspective of metabolite-oriented regulation does not contradict the perspective of pathway-oriented regulation; it expands upon it, by recognizing that many metabolites belongs to more than one pathway that tends to be co-regulated. If different needs in the metabolic network are met by co-regulation of different in-out pairs around a metabolite, but steady-state is maintained (*i.e.* the net change in production is always matched by changes in consumption, and *vice versa*),

then a metabolite's quantity is not strictly set by regulation of single pathway; instead, it emerges from variable network regulation. That is, the average of incoming fold changes will always be correlated with the average of outgoing fold changes, even if no single in-out pair is correlated all the time and some pairs are never correlated.

Such emergent regulation of the metabolite's quantity, distributed between various pathways, would be characterized by a correlation coefficient ρ between average producing reaction and average consuming reactions greater than the average ρ of all in-out pairs for that metabolite. Indeed, the ρ of averages was greater than the average genes expression correlation of in-out pairs for most metabolites (**Figure 4.3**B). In fact, in an idealized case, in which most pairs are correlated for a subset of the points, the ρ of average producing *vs.* the average consuming reactions would exceed the correlation coefficient of any in-out pair. Indeed, we found this to be true of 18 metabolites of the 145 that retained at least two in-out pairs after removal of reactions that are reversible and which input output was controlled by the same enzymes, and many more had correlations of the averages approaching that of the maximum in-out pair **Figure 4.3**A.

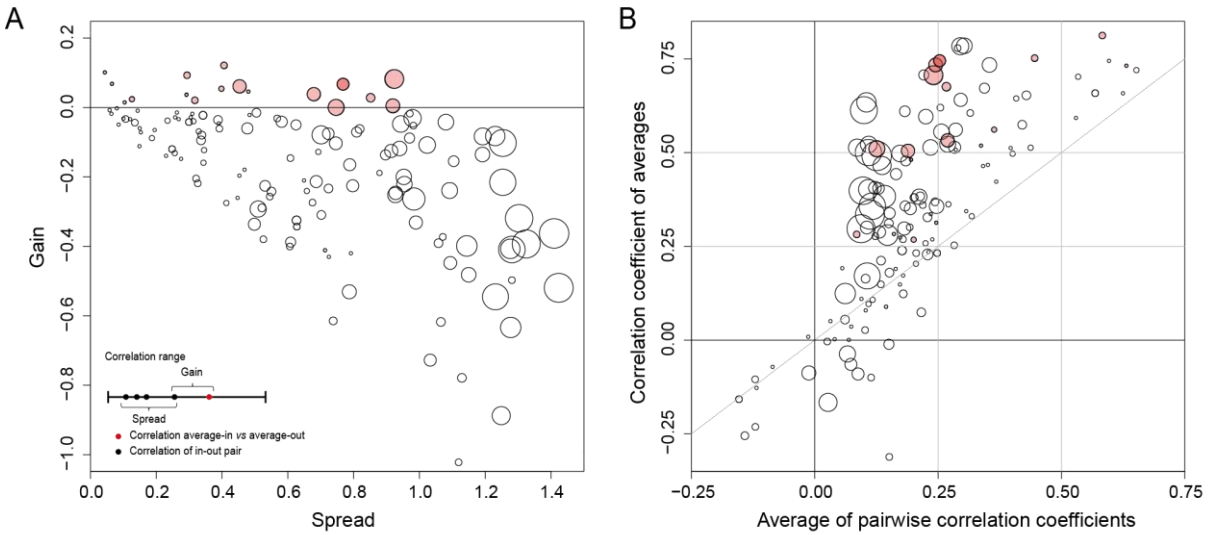


Figure 4.3 An emergent regulatory effect is found in a majority of metabolites. A) A collective action of all producing-consuming genes is higher than maximum individual pair (gain) for 13% of metabolites (red bubbles). The size of bubble is proportional to $2*\log_{10}$(number of in-out gene pairs). Metabolites having at least two in-out pairs are displayed accounting for a minimum of 3 genes. B) For majority of metabolites, correlation coefficients of average in-out are higher than average correlation of in-out pairs (above diagonal line). Red bubbles are the same as marked in panel A.

Our findings can be rationalized by the recognition that metabolites and their fluxes are the objects of metabolic regulation, and enzymes are merely the tools by which the cell manipulates concentration levels and fluxes. In most cases, a cell will have an interest in ensuring that

intermediate metabolite concentrations neither rise nor fall dramatically. In order to maintain such a balance, a metabolite's total incoming flux must match its total outgoing flux, but the particular route is unimportant from the perspective of maintaining the metabolite's concentration. The average gene expression fold change of the incoming reactions is correlated with the average fold change of the outgoing reactions in majority of metabolites. Such an emergent collective regulation of a metabolite's reactions is observed as a general co-expression pattern in metabolic network of *Saccharomyces cerevisiae* (**Figure 4.4**).

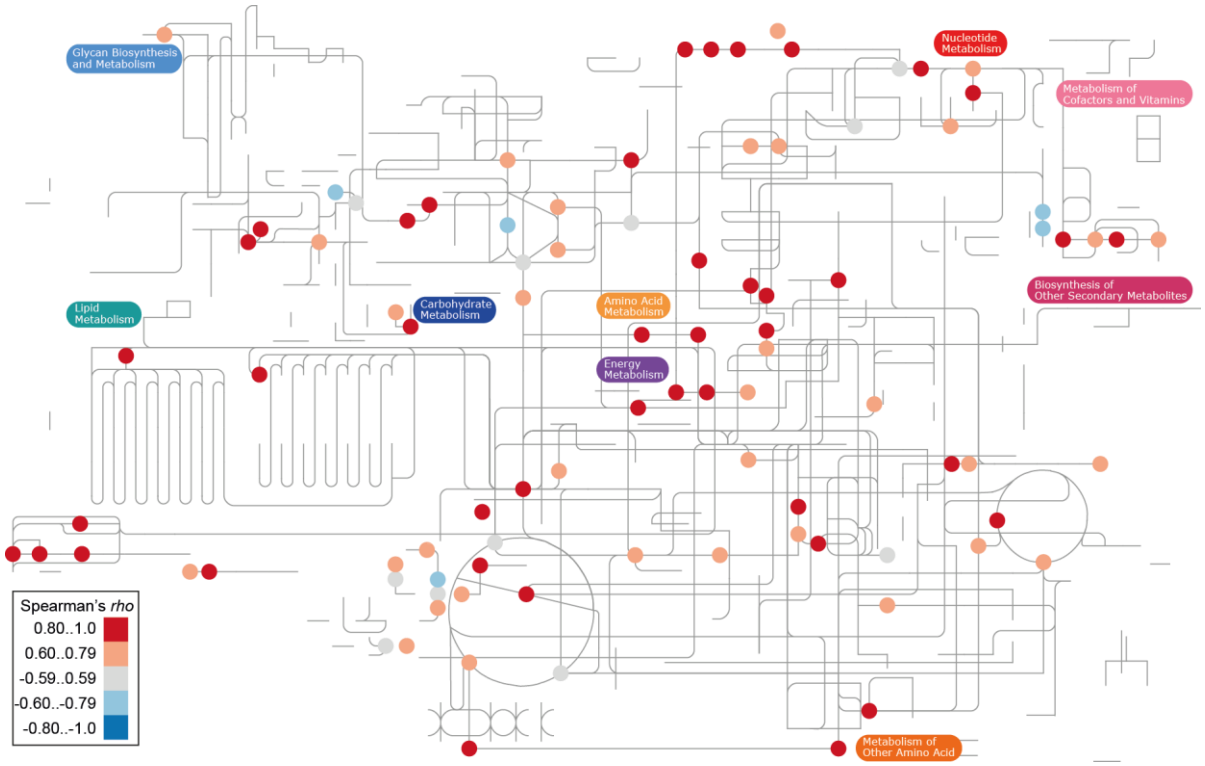


Figure 4.4 Emergent metabolite regulation in yeast metabolic network. Highlighted metabolites shows emergent co-regulation pattern. Figure was generated using: http://pathways.embl.de/

Materials and Methods

Datasets

ArrayExpress database was searched for term “glucose” and organism “Saccharomyces cerevisiae”, choosing only Affymetrix platform microarrays. Of the 69 datasets returned, we hand-selected only those which had a glycolytic carbon source, leaving 22. Each dataset consists of the data from a number of different microarray chips. Datasets were parsed to detect experimental factors, and then grouped the data from each chip based on these factors. All combinations of factors were tried, from grouping samples (chips) by all factors to grouping them by only one at a time. If multiple factors were used to group samples, only groups that differed in one factor were compared against each other, making in total 118 comparisons. Probes were annotated with the ‘annotate’ package from Bioconductor (Gentleman et al, 2004); For each comparison between groups, each probe’s signal was averaged among all the chips in the group, and fold change between the two groups’ averages for each probe was calculated.

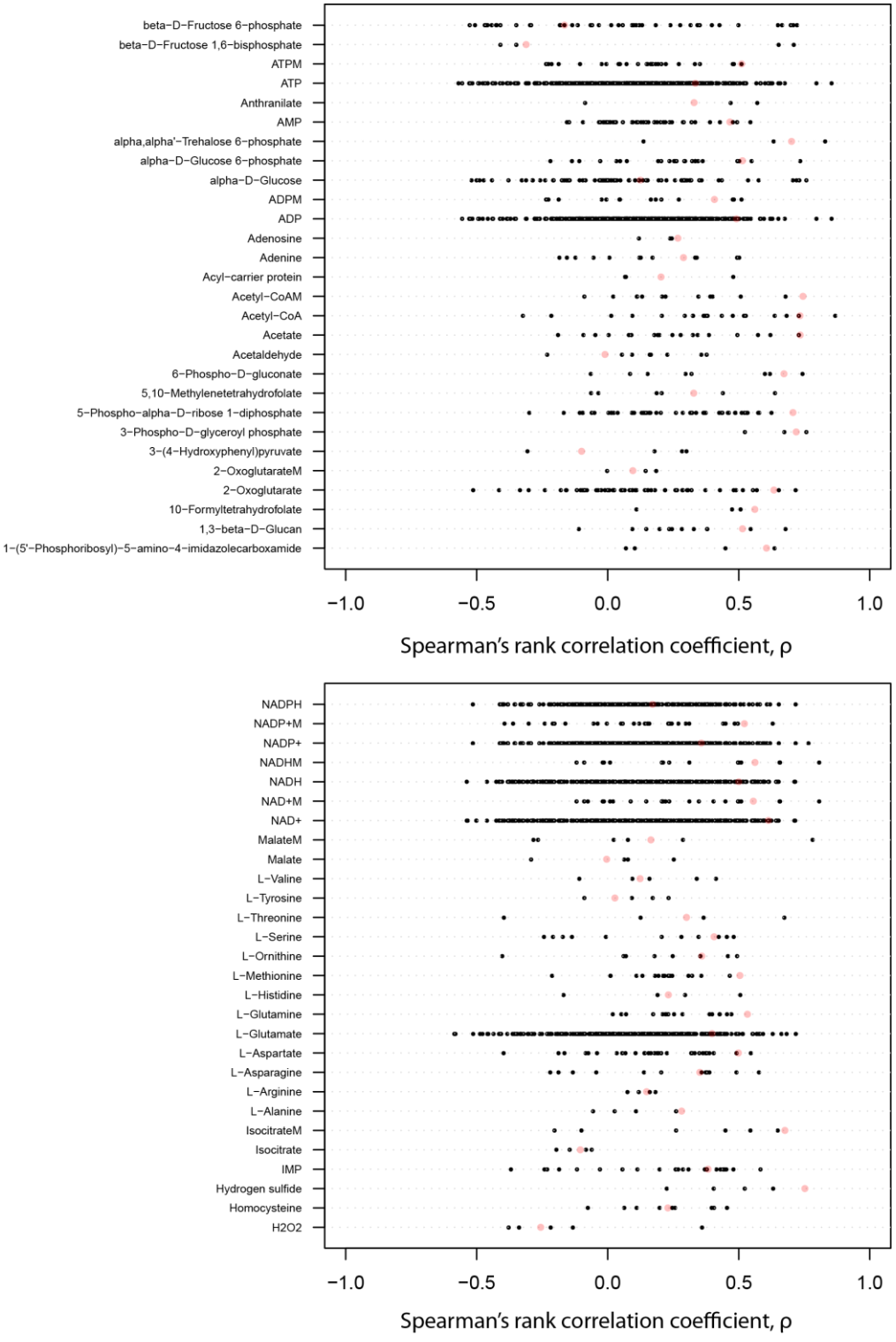
Metabolic network and flux variability analysis

Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network by (Forster et al, 2003) was used to link metabolites with the enzyme-coding genes. A total of 708 structural open reading frames (ORFs) are accounted for in this network, representing 1035 metabolic reactions (in total 1175 reactions and 584 metabolites). Reaction directions were estimated by using flux variability analysis (Mahadevan & Schilling, 2003). The steady-state model was constrained using glucose anaerobic growth physiological data (Fendt et al, 2010) (**Supplementary Table 4.1**). The resulting linear programming problem was solved using the glpk (http://www.gnu.org/software/glpk/) solver accessed through a C library. Reversible and non-feasible reactions were removed, leaving producing and consuming reactions. For reactions catalyzed by isoenzymes, we used the average fold changes of the enzyme-coding transcripts, while for complexes, we used the transcript with the lowest fold change. Applying these criteria, each reaction was assigned a single fold-change value (see **Chapter 3**, **Figure 3.1**B). For each pair of producing and consuming genes/reactions, the Spearman’s rank correlation coefficient ρ was calculated for the fold changes from each comparison. Additionally, for each comparison, we took the average of all the producing genes and the average for all the consuming genes and calculated the Spearman’s rank correlation coefficient. Calculations were performed using the statistical software R (www.r-project.org).

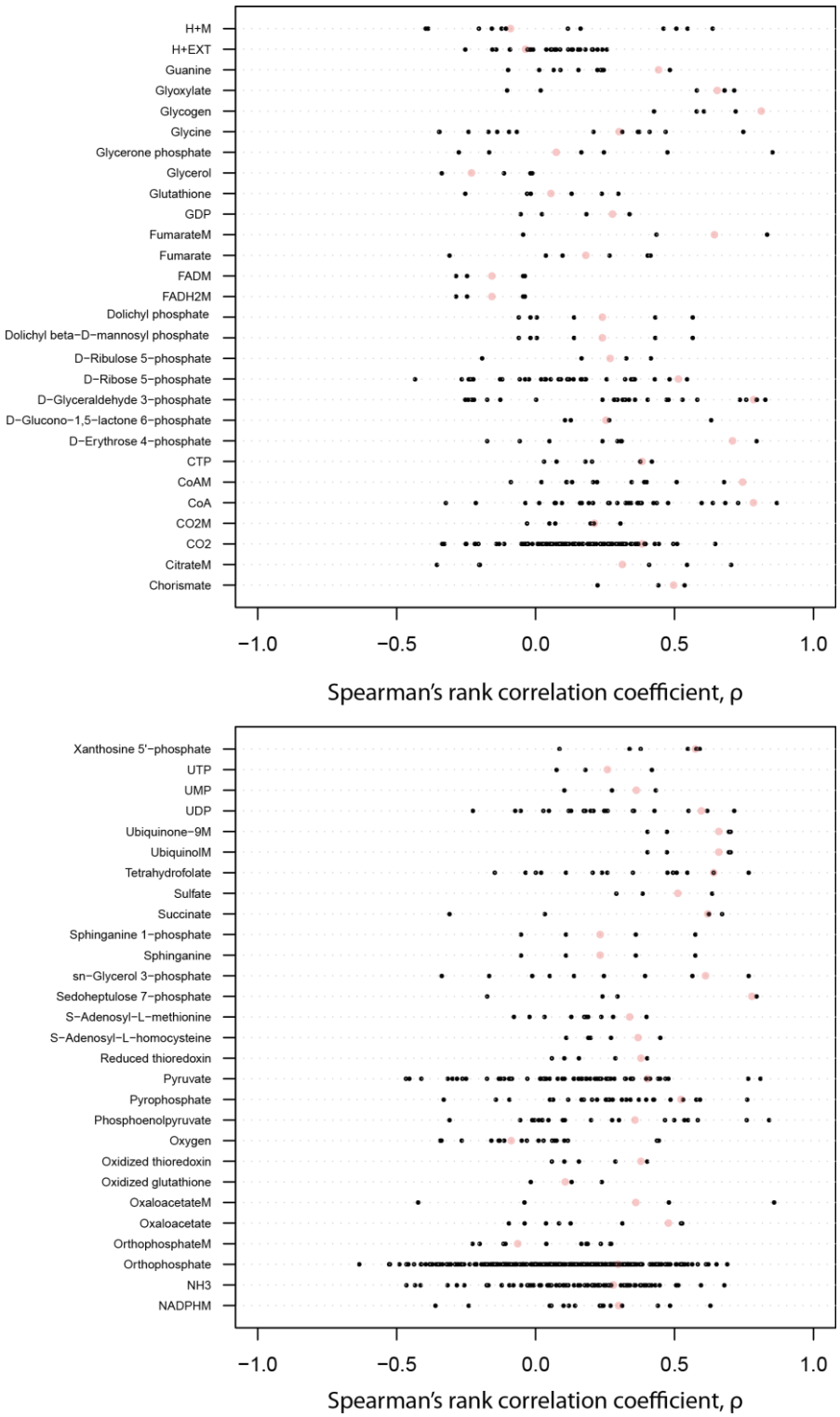
Supplementary Information

List of Supplementary Figures

List of Supplementary Tables



Supplementary Figure 4.1 Metabolite co-regulation patterns in yeast metabolic network. Panels are arbitrarily ordered, black dots represents individual input *vs.* output pair correlations; red dots represent the correlation of the average of input *vs.* average of output reactions.



Supplementary Figure 4.1 (continued) Metabolite co-regulation patterns in yeast metabolic network. Panels are arbitrarily ordered, black dots represents individual input *vs.* output pair correlations; red dots represent the correlation of the average of input *vs.* average of output reactions.

Supplementary Table 4.1 Physiological data of yeast batch anaerobic exponential growth (Fendt et al, 2010). Comma-separated values denote lower and upper bounds used for constraining the corresponding fluxes.

Reaction (mmol/g/h)	**WT**
Glucose uptake	16.15, 17.85
Ethanol secretion rate	20, 22
Glycerol secretion rate	2, 2.2
Acetate secretion rate	0.75, 1.51
Pyruvate secretion rate	0.06, 0.09
Growth rate	0.3, 0.33

Der Mensch ist, was er ißt.[*]

–Ludwig Andreas Feuerbach, 1863

[*] Translation from German: the man is what he eats

Chapter 5

Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes

Abstract*

Type 2 diabetes mellitus (T2DM) is a disorder characterized by both insulin resistance and impaired insulin secretion. Recent transcriptomics studies related to T2DM have revealed changes in expression of a large number of metabolic genes in a variety of tissues. Identification of the molecular mechanisms underlying these transcriptional changes and their impact on the cellular metabolic phenotype is a challenging task due to the complexity of transcriptional regulation and the highly interconnected nature of the metabolic network. In this study we integrate skeletal muscle gene expression datasets with human metabolic network reconstructions to identify key metabolic regulatory features of T2DM. These features include reporter metabolites – metabolites with significant collective transcriptional response in the associated enzyme-coding genes, and transcription factors with significant enrichment of binding sites in the promoter regions of these genes. In addition to metabolites from TCA cycle, oxidative phosphorylation, and lipid metabolism (known to be associated with T2DM), we identified several reporter metabolites representing novel biomarker candidates. For example, the highly connected metabolites NAD+/NADH and ATP/ADP were also identified as reporter metabolites that are potentially contributing to the widespread gene expression changes observed in T2DM. An algorithm based on the analysis of the promoter regions of the genes associated with reporter metabolites revealed a transcription factor regulatory network connecting several parts of metabolism. The identified transcription factors include members of the CREB, NRF1 and PPAR family, among others, and represent regulatory targets for further experimental analysis. Overall, our results provide a holistic picture of key metabolic and regulatory nodes potentially involved in the pathogenesis of T2DM.

* Published as: Zelezniak A, Pers TH, Soares SP, Patti ME, Patil KR. Metabolic Network Topology Reveals Transcriptional Regulatory Signatures of Type 2 Diabetes. PLoS Computational Biology, Vol. 6, No. 4, 2010, p. e1000729

Introduction

Type 2 diabetes mellitus (T2DM) is emerging as one of the main threats to human health in the 21st century with an estimated 300 million individuals with T2DM by the year 2025 (Simpson et al, 2003; Zimmet et al, 2001). T2DM is characterized by both insulin resistance, as manifested by reduced insulin-stimulated glucose uptake in skeletal muscle and adipose tissue and inappropriately high hepatic glucose output (Pehling et al, 1984; Shulman, 2000), and reduced insulin secretion by pancreatic β-cells (Muoio & Newgard, 2008; Shulman, 2000). Although the specific molecular pathophysiology remains unclear, many risk factors have been identified for T2DM, including family history of diabetes and prominent environmental factors such as alterations in early life development, excessive food intake, obesity, decreased physical activity and aging (Muoio & Newgard, 2008; Shulman, 2000; Simpson et al, 2003). At the cellular level, multiple regulatory mechanisms and metabolic pathways may contribute to the pathogenesis of insulin resistance, potentially mediated by alterations in insulin signaling (Saltiel & Kahn, 2001), mitochondrial oxidative metabolism and ATP production (Kelley et al, 2002; Mootha et al, 2003; Patti et al, 2003), fatty acid oxidation (Boden, 1996), or proinflammatory signaling (Ueki et al, 2004). Similarly, alterations in β-cell development and metabolism (Muoio & Newgard, 2008) may contribute to decreased insulin secretion.

Available human tissue transcriptome data related to T2DM (Sreekumar et al, 2002; Yang et al, 2002b) provide an opportunity for identification of novel molecular mechanisms underlying the metabolic phenotype of T2DM. This task is challenging due to the need to account for the inherent high connectivity of bio-molecular interaction networks. We have utilized a network-centered methodology to link diabetes-related alterations in gene expression to metabolic hot spots and transcription factors potentially responsible for gene expression changes.

Rationale and Methodology

Metabolic phenotypes at a cellular level are essentially characterized by concentrations of metabolites and fluxes through the reactions that make up the metabolic network. Fluxes, in turn, are dependent on metabolite levels, enzyme activities, abundance of effectors and possibly other variables. Measurement of fluxes and metabolite concentrations at the entire metabolic network-scale is, however, a difficult task in humans due to a variety of technological and experimental limitations. By contrast, methods for measurement of expression of genes encoding metabolic enzymes are relatively well-established. Thus, the primary goal of this study is to use informatics

approaches to integrate available gene expression data with metabolic networks, in order to predict metabolic phenotypes of skeletal muscle linked to the pathogenesis of type 2 diabetes. Such an approach will help not only to gain insight into the organization of transcriptional regulation in human tissues, but also provide guidance for improved design of experimental strategies for obtaining metabolite and flux data, which can be further integrated into metabolic models.

To achieve these goals, we applied an extension of the algorithm described in (Patil & Nielsen, 2005) (for various applications of this algorithm see (Baxter et al, 2007; Capel et al, 2009; David et al, 2006; Patil & Nielsen, 2005; Seggewiss et al, 2006)), which enables identification of so-called reporter metabolites, or metabolic hot spots around which transcriptional regulation is centered (**Figure 5.1**A, Supporting text S). This analysis is based on the assumption that under most conditions of physiological interest, fluxes through enzymes connected to a metabolite are coordinated in order to maintain physiological homeostasis, or to eventually reach a new (pseudo-) steady-state. Moreover, transcriptional regulation of expression of genes encoding critical enzymes in metabolic flux pathways facilitates concordance with the metabolic demands of the cell and corresponding stoichiometric and thermodynamic constraints on fluxes. For this analysis, we applied two recently published human metabolic network models: i) *Homo sapiens* Recon1 (Duarte et al, 2007), and ii) Edinburgh Human Metabolic Network (EHMN) (Ma et al, 2007).

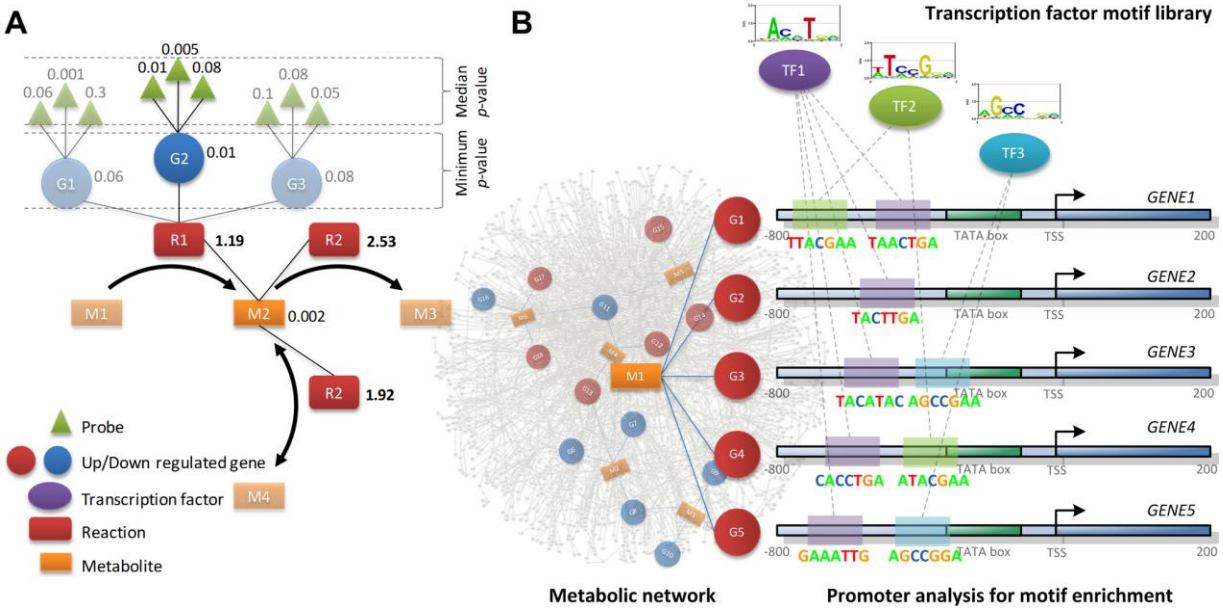


Figure 5.1 Schematic overview of the methodology used for the identification of reporter metabolites and associated putative regulatory sequence motifs. A) Scoring system for identification of reporter metabolites. Each metabolite is scored based on the scores of the associated enzyme-catalyzed reactions. Each enzyme, in turn, is assigned a score based on median of the *P*-values of the probes representing the corresponding gene. In case of a reaction catalyzed by an enzyme complex or a set of isozymes, minimum of the *P*-values of the corresponding enzymes is chosen. Numbers in bold are Z-scores for each reaction, the rest of the numbers represent *P*-values (significance of differential expression). B) Identification of transcription factor binding motifs. For a reporter metabolite, a set of up/down regulated neighbor (enzyme-coding) genes is selected. Promoter regions, upstream of transcription start site (TSS) of each of the selected genes are assessed for the enrichment of known transcription factor (TF) binding sequence motifs.

We further hypothesized that the observed coordinated changes around reporter metabolites can be, at least in some cases, attributed to common transcriptional regulatory mechanisms. Specifically, we hypothesize that the neighbor enzymes of reporter metabolites may share one or more transcription factor binding sites in the promoter regions of the corresponding genes. In order to identify such potential regulatory players, we tested promoter sequences of the genes associated with the reporter metabolites for enrichment of known transcription factor binding motifs (**Figure 5.1**B). Transcription factors identified in this fashion provide clues to the regulatory mechanisms that lead to observed gene expression changes in the metabolic network.

Since our goal is to identify reporter metabolites and transcription factors potentially involved in diabetes pathogenesis and progression we analyzed two independent studies of skeletal muscle transcriptomics in individuals with established type 2 diabetes or insulin resistance (Mootha et al, 2003; Patti et al, 2003) (**Supporting Text 1**), In the first study (Mootha et al, 2003), biopsies were obtained following insulin stimulation from a cohort of 43 Swedish men of Caucasian ethnicity with a spectrum of glucose tolerance, including 17 with normal glucose tolerance (NGT), 8 with impaired glucose tolerance (IGT), and 18 with established T2DM. The second dataset (Patti et al, 2003) was

derived from a cohort of 15 subjects of Mexican American ethnicity, in whom muscle biopsies were performed in the fasting state. Importantly, this cohort included individuals with not only established diabetes (5 subjects, T2DM), but also individuals with completely normal glucose tolerance but a spectrum of insulin resistance; normal glucose tolerant subjects were subdivided by family history-linked diabetes risk (4 family history positive, more insulin resistant subjects, FH+; and 6 family history negative, more insulin sensitive subjects, FH-). With this approach, the individual contributions of isolated insulin resistance and diabetes risk (in the setting of normoglycemia, FH+), mild elevations in postprandial glucose (IGT), and established diabetes can be individually assessed. Moreover, the possible contribution of family history, potentially mediated by genetics or shared environment, can be assessed. Thus, we predict that analysis of the common patterns resulting from the two datasets will identify regulatory signatures potentially independent of study cohort and design variation but common to the pathophysiology of insulin resistance and diabetes.

Results

In present study, we performed reporter metabolite analysis based on pair-wise comparisons within each dataset; differential expression and its significance were assessed with robust multi-array average (RMA) and empirical Bayes testing. Significance of differential expression for each gene was used as a scoring metric (Materials and Methods). The results are summarized as metabolic signatures (reporter metabolites) and regulatory signatures (transcription factors) for T2DM.

Metabolic signatures of T2DM

Swedish male dataset

Reporter metabolite analysis for three pair-wise comparisons, *viz.,* T2DM *vs.* NGT, T2DM *vs.* IGT, and IGT *vs.* NGT, revealed significant reporter metabolites (*P*-value ≤ 0.05) participating in lipid metabolism, TCA cycle, oxidative phosphorylation (OXPHOS) and glycolysis (**Table 5.1**, **Table 5.2**, Table S1[†] and Table S2). Among reporter metabolites identified for the T2DM *vs.* NGT comparison were lipid species 1,2-diacyl-sn-glycerol (DAG), acetoacetyl-CoA, and the sphingolipid sphinganine. These are interesting as prior studies (Holland et al, 2007; Roden, 2005; Savage et al, 2007; Shulman, 2000) have demonstrated that the related lipid molecules diacylglycerols (DAG), long-chain fatty acyl CoAs, and ceramides correlate positively with triglyceride content and inversely with insulin sensitivity (Muoio & Newgard, 2008) and have been shown to induce insulin resistance (Shulman,

[†] Supplementary tables S1-S8 together with Figure S1 for this chapter can be accessed through the web: http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000729#s5

2000). Furthermore, given that saturated fatty acids appear to play a particularly important pathogenic role in insulin resistance (Salek et al, 2007), it is interesting that several metabolites of saturated fatty acids (such as hexanoyl-CoA, palmitoyl-CoA, tetradecanoyl-CoA, lauroyl-CoA, decanoyl-CoA and butanoyl-CoA) were found as reporter metabolites with mostly up-regulated neighboring genes in the IGT *vs.* NGT comparison (**Table 5.1** and Table S2), and thus may serve as potential markers of insulin resistance and IGT.

TCA cycle metabolites citrate and 2-oxoglutarate, with down-regulated neighboring genes, were also uncovered as reporter metabolites in the T2DM *vs.* NGT comparison (**Table 5.1**, Table S1, Table S2). These results are concordant with a study of human urine metabolome profiles from patients with T2DM (Salek *et al.*, 2007), in which levels of citrate and 2-oxoglutarate were lower in T2DM compared to healthy controls (Newgard et al, 2009). Among other mitochondrial metabolites, reduced and oxidized forms of cytochrome c and ubiquinol were identified as reporter metabolites (T2DM *vs.* NGT, Table S1) with down-regulated expression of the associated genes.

Impaired glucose tolerance typically reflects an important transition between normoglycemia and overt diabetes, reporter metabolites which are identified in both IGT *vs.* NGT and T2DM *vs.* NGT, but not significantly different in the T2DM *vs.* IGT comparison (*e.g.* phosphatidylethanolamine, 2-hydroxyglutarate, 2-oxoglutarate, 3',5'-cyclic AMP, ATP, Table S1, Table S2) may be considered novel biomarkers of early-stage glucose intolerance.

Mexican-American dataset

We similarly performed reporter metabolite analysis using both Recon1 and EHMN metabolic models in the Mexican-American dataset. This analysis revealed significant transcriptional regulation in metabolite nodes in TCA cycle, oxidative phosphorylation, and lipid metabolism, for both T2DM *vs.* FH- and FH+ *vs.* FH- comparisons (Patti et al, 2003). Similar to the Swedish Caucasian dataset, metabolites involved in oxidative phosphorylation (*e.g.* ferrocytochrome c, H+, and fumarate) were among the top-ranking reporter metabolites, identified in both the T2DM *vs.* FH- and FH+ *vs.* FH- comparisons (**Table 5.2**, Table S3). Interestingly, urinary levels of fumarate, an important link between the TCA cycle and oxidative phosphorylation, were recently found to be decreased in T2DM patients (Salek et al, 2007).

Analysis using the EHMN model revealed TCA cycle-related metabolites, including 3-carboxy-1-hydroxypropyl-ThPP, aconitate, succinyl-CoA, malate and fumarate, as significant reporter metabolites (P-value ≤ 0.05), with mostly down-regulated expression of the genes encoding their

neighboring enzymes. Ubiquinol was found as reporter metabolite representative of electron transfer chain. Several molecules within β-oxidation pathways, such as 3-cis-dodecenoyl-CoA, glutaryl-CoA, trans-3-decenoyl-CoA, 3-methylbutanoyl-CoA and 3-methylcrotonyl-CoA, as well as in amino acid (leucine, lysine) metabolism were also identified as reporters (**Table 5.2**, Table S4). Moreover, glutamate, glycerol derivatives, phosphocreatine, a number of hormone derivatives and many others (Table S3 and Table S4) were found as significant reporter metabolites in the T2DM *vs.* FH- comparison.

Table 5.1 Reporter metabolites for Swedish male dataset.

Reporter Metabolite	P-values		Enzyme neighbors (Up-regulated : Down-regulated)		
	T2DM/NGT	**IGT/NGT**	**T2DM/NGT**	**IGT/NGT**	
Citrate	**0.047**	0.646	1:0	1:0	TCA cycle
Succinyl-CoA	**0.013**	0.285	2:3	2:3	
2-Hydroxyglutarate*	**0.002**	**0.023**	0:1	0:1	
*2-Oxoglutarate**	**0.049**	**0.047**	8:11	8:11	
Ferrocytochrome C; Ferricytochrome C	**0.006**	**0.032**	1:2	0:3	OXPHOS
Ubiquinone-10	**0.017**	0.769	0:5	1:4	
Ubiquinol-10	**0.022**	0.484	0:4	1:3	
Phosphoenolpyruvate*	0.196	**0.037**	1:3	1:3	Glycolysis
D-Glyceraldehyde*	0.083	**0.017**	2:1	3:0	
D-Alanine	**0.016**	0.330	0:3	0:3	A. a. metabolism
L-Alanine	**0.047**	0.319	3:7	3:7	
3-Methylglutaconyl-CoA†	**0.038**	0.816	0:2	1:1	
*L-Leucine**	**0.047**	0.109	1:3	1:3	
*1,2-Diacyl-sn-glycerol (DAG)**	**0.022**	**0.049**	2:5	2:5	Lipid metabolism
1D-myo-Inositol 1,4-bisphosphate†	0.060	0.151	0:3	2:1	
3-Dehydrosphinganine*	0.232	**0.035**	1:1	2:0	
Acetoacetyl-CoA*	**0.009**	0.462	1:4	2:3	
Butanoyl-CoA†	0.365	**0.038**	0:2	1:1	
*Decanoyl-CoA; Lauroyl-CoA**	0.268	**0.033**	1:2	2:1	
Fatty acid*	**0.021**	0.756	3:4	3:4	
Lophenol*§	**0.007**	0.749	0:1	0:1	
*Palmitoleoyl-CoA**	0.238	**0.019**	1:3	2:2	
*Palmitoyl-CoA**	0.179	**0.014**	3:4	6:1	
Phosphatidyl glycerol phosphate	**0.047**	0.316	0:1	0:1	
Phosphatidylinositol 4,5-bisphosphate	0.097	**0.001**	1:5	2:4	
Propanoyl-CoA*	0.259	**0.016**	2:5	2:5	
Prostaglandin E2	**0.036**	**0.032**	0:3	1:2	
Sphinganine*	**0.038**	0.283	1:3	2:2	
(Gal)3 (GalNAc)1 (Glc)1 (Cer)1*	**0.023**	**0.034**	1:2	1:2	Other
AMP†	**0.041**	0.218	7:17	6:17	
ATP†	**0.003**	**0.010**	28:60	27:60	
cAMP†	**0.033**	**0.049**	2:0	2:0	
CDPcholine	**0.020**	0.122	0:2	0:2	
Choline phosphate	**0.030**	0.573	0:2	1:1	
NAD+*	0.333	**0.020**	29:34	34:34	
Phosphocreatine	**0.025**	0.176	0:1	1:0	
Trichloroethanol*	**0.020**	**0.038**	1:2	3:0	

*Reporter metabolites identified using EHMN metabolic network. †Reporter metabolites identified in both networks. §Plant metabolite, likely to be present in the EHMN due to incorrect annotation. Reporter metabolites with $p \leq 0.05$ in at least one of the comparisons showed in bold. Columns with enzyme neighbors show the number of up- and down-regulated enzyme neighbors in the first condition (*e.g.* T2DM/NGT up- and down-regulated in T2DM comparing with NGT) for each of comparisons. Reporter metabolites without marks were identified using Recon1 metabolic network. Metabolites written in italics are known to be directly/indirectly related to T2DM, see main text and Table S8. A.a. – amino acid; OXPHOS – oxidative phosphorylation.

Table 5.2 Reporter metabolites for Mexican-American dataset.

Reporter metabolite	P-values		Enzyme neighbors (Up-regulated : Down-regulated)		
	T2DM/FH-	FH+/FH-	T2DM/FH-	FH+/FH-	
2-Oxoglutarate	**0.001**	**0.001**	2:7	2:7	TCA cycle
L-Malate	0.098	**0.029**	1:4	2:3	
Succinyl-CoA[†]	**0.011**	**0.009**	0:5	0:5	
Ferrocytochrome C;Ferricytochrome C	**0.008**	**0.007**	0:3	0:3	OXPHOS
Fumarate	**0.019**	**0.025**	0:2	0:2	
Ubiquinone-10[†];Ubiquinol-10[†]	**0.040**	**0.021**	1:3	1:3	
2,3-Disphospho-D-glycerate[†]	**0.021**	**0.004**	0:1	0:1	Glycolysis
2-Phospho-D-glycerate[*]	**0.038**	**0.006**	0:2	1:1	
beta-D-Fructose[*]	**0.049**	**0.038**	0:2	0:2	
D-Fructose 2,6-bisphosphate	**0.037**	0.136	0:2	0:1	
D-Fructose 6-phosphate	**0.013**	0.119	4:6	3:7	
D-Glucose[*]	**0.037**	0.066	0:7	1:5	
D-Glucose 6-phosphate	**0.009**	**0.014**	1:3	1:3	
D-Glycerate 2-phosphate	**0.026**	**0.003**	0:2	1:1	
L-Lactate	**0.048**	0.067	1:2	1:2	
Phosphoenolpyruvate	0.079	**0.048**	2:2	3:1	
Pyruvate	**0.042**	0.202	1:6	1:6	
2-Oxoadipate[*]	**0.002**	**0.004**	0:1	0:1	A. a. metabolism
beta-Alanine	**0.031**	**0.027**	1:1	1:1	
L-Glutamate[†]	**0.025**	**0.009**	1:1	1:1	
(R)-2-Methyl-3-oxopropanoyl-CoA[*]	**0.043**	0.118	0:2	0:1	Lipid metabolism
1,2-Diacyl-sn-glycerol (DAG)[*]	**0.036**	0.117	3:2	5:1	
1D-myo-Inositol 1,4-bisphosphate	**0.025**	0.054	1:2	1:2	
3-cis-Dodecenoyl-CoA[*]	**0.009**	**0.039**	0:3	0:3	
Acylglycerol[*]; 2-Acylglycerol[*]	**0.035**	**0.018**	1:1	1:1	
Glutaryl-CoA[†]	**0.007**	**0.015**	0:2	0:2	
Glycerol	**0.020**	**0.001**	1:1	1:1	
Glycerol 3-phosphate	0.051	**0.005**	2:1	2:1	
Lipoamide[*]	**0.014**	**0.006**	0:5	0:5	
Phosphatidylinositol	**0.017**	0.128	1:5	1:5	
trans-3-decenoyl-CoA[*]	**0.026**	0.076	0:2	0:2	
ADP	**0.047**	0.174	16:31	20:27	Other
CO_2	**0.041**	**0.004**	1:11	3:9	
Coenzyme A[†]	**0.007**	**0.014**	4:8	3 10	
Creatine;Phosphocreatine[†]	**0.032**	**0.048**	0:1	0:1	
NAD+[†]; NADH[†]	**0.003**	0.095	3:17	17:4	
Trichloroethanol[*]	**0.021**	**0.006**	2:1	3:0	

[*]Reporter metabolites identified using EHMN metabolic network. [†]Reporter metabolites identified in both networks. Reporter metabolites with $p \leq 0.05$ in at least one of the comparisons showed in bold. Columns with enzyme neighbors show the number of up- and down-regulated enzyme neighbors in the first condition (*e.g.* T2DM/FH- up- and down-regulated in T2DM comparing with FH-). Reporter metabolites without marks were identified using Recon1 metabolic network. Metabolites written in italics are known to be directly/indirectly related to T2DM, see main text and Table S8. A.a. – amino acid; OXPHOS – oxidative phosphorylation.

Overlapping reporter metabolites between two study populations

In order to determine the extent of overlap between the two study populations, we performed a cluster analysis of the pair-wise comparisons within the Swedish and Mexican-American datasets (**Figure 5.2**). Jaccard distance metric between two pair-wise comparisons (*e.g.* T2DM *vs.* FH- and FH+ *vs.* FH-) was calculated based on the overlap of reporter metabolites between the two comparisons. Jaccard distance provides a measure of dissimilarity between two sets of reporter metabolites, and is quantified as the fraction of non-overlapping reporter metabolites between the two sets. While similar clustering patterns were observed (**Figure 5.2**A and **Figure S1**A) independent of the use of either EHMN or Recon1 metabolic model, Swedish and Mexican-American studies clustered separately, perhaps related to differences in study population, study design (*e.g.* fasting studies in Mexican-Americans, insulin-stimulated studies in Swedish) or differences in microarrays used (thus differing in the coverage of metabolic enzymes). We observed substantial overlap between the T2DM *vs.* FH- and FH+ *vs.* FH- comparisons, suggesting that insulin resistance patterns could contribute to these findings (Table S5, Table S6).

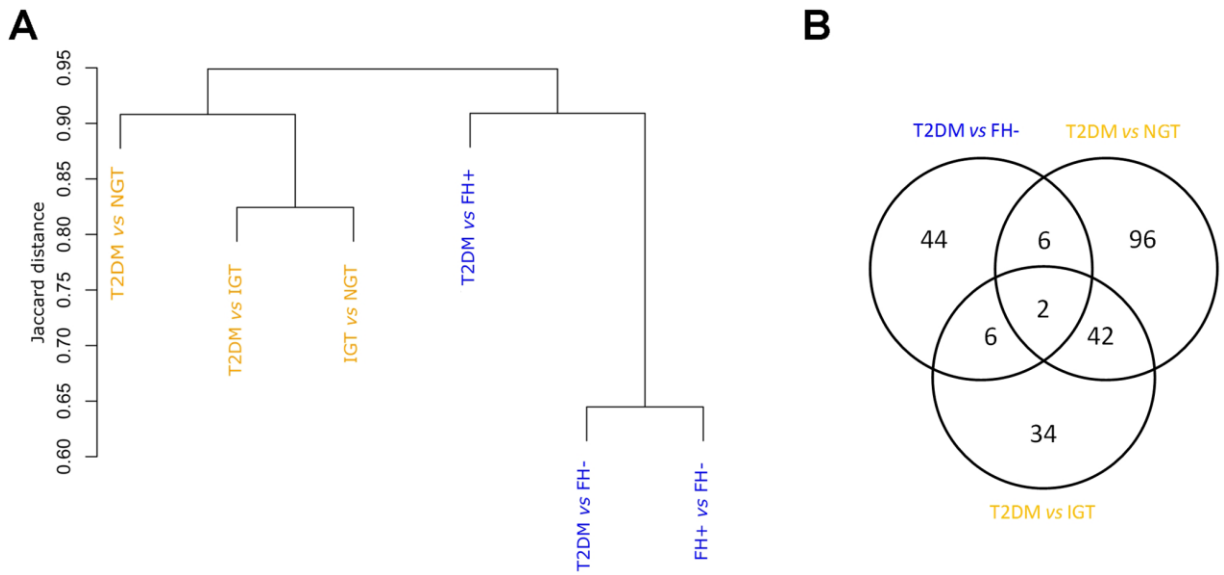


Figure 5.2 Hierarchical clustering of pair-wise comparisons within the Swedish male and Mexican-American datasets based on the overlapping reporter metabolites (Recon1 model). Comparisons are colored according to the dataset; blue – Mexican-American; orange – Swedish male dataset. A) Dendrogram of reporter metabolites identified in each of the comparisons based on Jaccard distance. B) Venn diagram showing the overlap of the reporter metabolites identified in the different comparisons.

We next examined the overlap of reporter metabolites between the two case studies (**Figure 5.2**B and **Figure S1**B). Owing to differences in the metabolite-gene connectivity between EHMN and Recon1, the number of overlapping reporter metabolites is generally higher for the EHMN analysis. To a large extent, this difference is due to the groups of metabolites in EHMN that share the same

gene neighbors (whether two metabolites share the same gene neighbors depends not only on the network used, *i.e.* number of distinct biochemical reactions associated with a particular enzyme, but also on the coverage of genes on the particular microarray chip used). In addition to many other metabolites, phosphocreatine appeared as a significant reporter in both case studies, *viz.*, for T2DM *vs.* NGT and T2DM *vs.* FH- comparisons. Phosphocreatine is an important energy reservoir metabolite in skeletal muscle, and defects in recovery of phosphocreatine have been identified *in vivo* in humans with insulin resistance (Fleischman et al, 2009) and diabetes (Phielix et al, 2008). Interestingly, low levels of urinary creatine have also been found in patients with T2DM (Salek et al, 2007).

Regulatory signatures of T2DM

In order to link the identified reporter metabolites to regulatory pathways controlling gene expression, we hypothesized that enzymes associated with reporter metabolites would be regulated by common transcription factors. As potential candidates subjected to such regulation we selected all reporter metabolites with at least 5 up- or down-regulated neighboring genes (Materials and Methods). Up- and down-regulated gene sets were then analyzed separately in order to assess whether their promoter regions were enriched for known transcription factor binding sequence motifs. *P*-values for enrichment were estimated by using a hyper-geometric test, which compared the proportion of promoters from a given gene set containing a particular motif with the frequency of occurrence of that motif in promoter regions of all other metabolic genes. Correction for multiple-testing was done by using q-value (Storey & Tibshirani, 2003) and motifs with q-value ≤ 0.05 were considered as significantly enriched.

In accord with our hypothesis, several transcription factor binding sites were overrepresented in the promoter regions of the enzymes associated with reporter metabolites. A summary of the main results from this analysis is illustrated in **Figure 5.3**A. Many transcription factors were found to be common across the two case studies (**Figure 5.3**B), *albeit* in connection with different reporter metabolites. PPAR family motifs (PPARγ and PPARα:RXRα) were enriched in seven downregulated enzyme sets including ATP. Tax/CREB motifs were enriched in promoters of downregulated enzymes associated with ATP, ADP and phosphate. Additional down-regulated neighbors of ATP were enriched for the binding sites of NF-κB, MEF-2, UF1-H3β, Pax-9 and NKX6.2, while the NRF-1 motif was enriched in the set of up-regulated enzymes neighboring ADP. Another potential regulatory signature was identified around the down-regulated neighbors of phosphatidylinositol and phosphatidylinositol 4,5-bisphospate (important phospholipids which participate in insulin and other signaling reactions), which were significantly enriched for binding sites of p53, PPARγ, SRF, SEF-1, v-

Jun, GCNF, AR and many others (Table S7). These and other highly connected reporter metabolites in the metabolite-TF network (**Figure 5.3**A) demonstrate the concept that associated metabolic pathways can be transcriptionally regulated in multiple ways in response to environmental stimuli or metabolic perturbation.

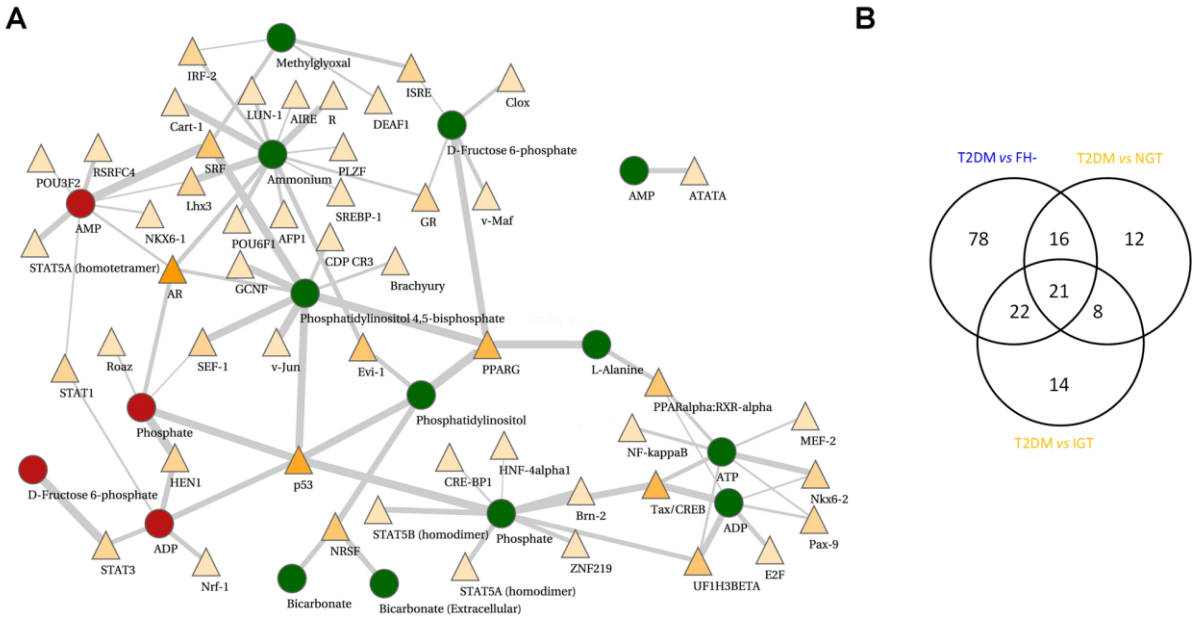


Figure 5.3 Summary of the main results from the motif enrichment analysis. A) Motif enrichment analysis for the genes associated with reporter metabolites from the T2DM *vs.* NGT comparison. Reporter metabolites with up-regulated neighboring gene set are shown as red circles, whereas reporter metabolites with down-regulated neighboring gene set are represented as green circles. Transcription factor binding motifs (shown as triangles) are colored according to the number of enzyme sets in which they are enriched, ranging from light yellow (enriched in few sets) to orange (enriched in as many as 6 sets). Edges are scaled according to q-values signifying the confidence of the motif enrichment. B) Venn diagram showing the overlap of transcription factor binding motifs across the comparisons of T2DM with non-T2DM cases. Comparisons are colored according to the dataset; blue – Mexican-American; orange – Swedish male dataset.

Discussion

Maintenance of whole-body glucose metabolism is reliant on a delicately balanced dynamic interaction between tissue sensitivity to insulin (including muscle, adipose and liver) and insulin secretion (Bajaj & Defronzo; Muoio & Newgard, 2008). Unfortunately, the molecular mechanisms responsible for diabetes risk remain unknown. A key metabolic phenotype associated with insulin resistance in humans is inappropriate lipid accumulation in tissues outside of adipose tissue, suggesting defects in fatty acid uptake, synthesis, and/or oxidation. With lipid excess and/or impaired oxidation, as observed in obesity and/or inactivity, flux of long-chain acyl CoAs (LC-CoA) may be redirected into cytosolic lipid species such as diacylglycerols (DAG), triacylglycerols (TG) and ceramides (derivatives of sphingosine and fatty acid metabolism) (Muoio & Newgard, 2008) that are

correlated with reductions in insulin signaling and insulin resistance (Holland et al, 2007; Itani et al, 2002; Roden, 2005; Savage et al, 2007; Shulman, 2000). Whether alterations in mitochondrial oxidative function in humans with insulin resistance and diabetes contribute to, or are a consequence of these defects, remains unclear (Patti & Corvera, 2010).

Recognizing these important gaps in our knowledge of diabetes pathophysiology, we have integrated transcriptomic data with metabolic networks to systematically identify, in an unbiased fashion, regulatory hot spots (reporter metabolites and associated transcription factors) associated with insulin resistance and T2DM. Our reporter metabolite results provide evidence for transcriptional dysregulation of multiple metabolic pathways in skeletal muscle. Interestingly, many of the reporter metabolites identified in our analysis have been appreciated in prior experimental studies in animal models (metabolites with italic font in **Table 5.1**, **Table 5.2** and Table S8). A bird's-eye view of selected metabolic and regulatory nodes identified in our study, integrated with selected key cellular pathways, is depicted in **Figure 5.4**, and key regulatory nodes potentially contributing to diabetes-associated pathophysiology are discussed in more detail below.

Key metabolic regulatory nodes in T2DM pathogenesis

Lipid metabolism

In conditions of overnutrition and physical inactivity, availability of cellular fatty acids stimulate ligand–dependent PPARα/δ transcription factors which, in turn, induce transcription of genes responsible for β-oxidation (Kersten et al, 2000; Koves et al, 2005). Metabolic byproducts of incomplete β-oxidation, such as acylcarnitines and reactive oxygen species, may accumulate in mitochondria and contribute to insulin resistance (Muoio & Newgard, 2008). Interestingly, our analysis identified enrichment of PPAR family transcription factor binding motifs in T2DM as compared with insulin sensitive subjects, in both the Swedish and Mexican-American datasets (T2DM *vs.* NGT and T2DM *vs.* FH-, respectively). Moreover, reporter analysis revealed lipid metabolites (Table S1), known to be natural ligands of PPARγ (prostaglandins) (Kersten et al, 2000).

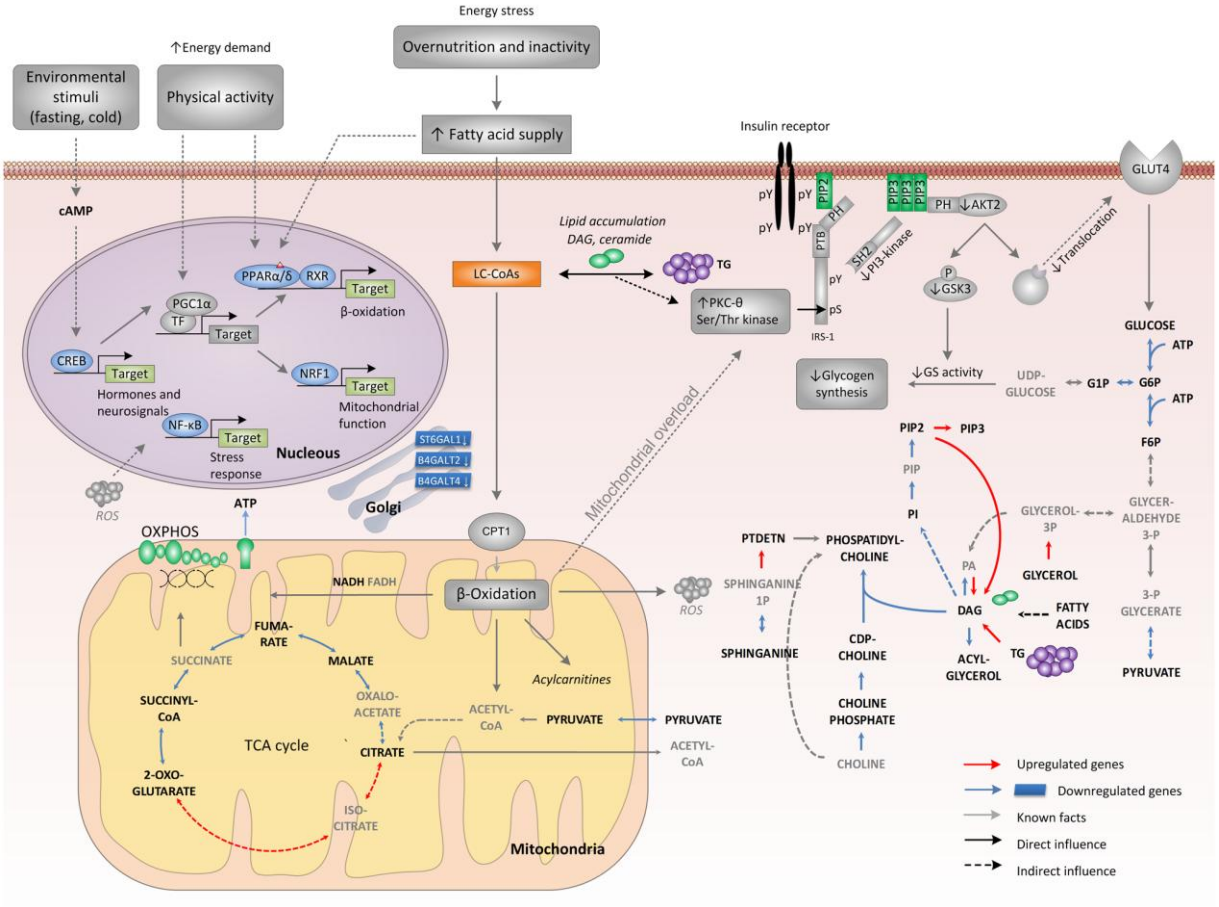


Figure 5.4 Metabolic and regulatory signatures of type 2 diabetes. Key metabolic and regulatory pathways associated with reporter metabolites identified in this study (T2DM *vs.* NGT and T2DM *vs.* FH- comparisons) are shown. Metabolites in bold black font are reporter metabolites. Grey shapes and arrows represent facts/hypotheses from previous studies and are not directly based on the results from the present study. Broken lines imply indirect effect while full lines denote direct effect. Chronic overfeeding and physical inactivity increase the influx of fatty acid, which promotes β-oxidation through the activation of PPARα/δ-mediated genes, without coordinated increase in TCA cycle flux. Reporter analysis supports this idea by showing the decreased activity in TCA cycle enzymes associated with reporter metabolites. Eventually, this leads to mitochondrial accumulation of metabolic by-products of incomplete β-oxidation (acylcarnitines ROS). These stresses might lead to mitochondrial overload which together with intracellular lipid-signaling (such as DAG) molecules might trigger serine a serine/threonine (Ser/Thr) kinase (Ser/Thr) cascade initiated by nPKCs. As a result, Ser/Thr phosphorylation of insulin receptor substrate 1 (IRS-1) sites is induced, thereby inhibiting IRS-1 tyrosine phosphorylation and activation of PI 3-kinase, resulting in impeded GLUT4 translocation, reduced glucose transpor, and decreased glycogen synthesis. Increased physical activity/fasting activates PGC1α and CREB (a potent inducer of PGC-1α). These actions combat lipid stress by increasing TCA cycle flux and by coupling ligand-induced PPARα/δ activity with PGC1α-mediated remodeling of downstream metabolic pathways such as respiration and β-oxidation. CDP-choline, cytidine diphosphate choline; DAG, diacylglycerol; G1P, glucose 1-phosphate; G6P, glucose 6-phosphate; GLUT4, glucose transporter-4; GSK3, glycogen synthase kinase-3; IRE1, inositol requiring kinase-1; LC-CoAs, long-chain acyl CoAs; nPKCs, novel protein kinase Cs; PA, phosphatidate; PGC1α, PPARγ co-activator-1α; PH, pleckstrin homology domain;PI, phospatidylinositol; PIP, phospatidylinositol 4-phospate; PIP2, phosphatidylinositol 4,5-bisphospate, PIP3, phospatidylinositol 3,4,5-trisphospate; PI 3-kinase, phosphoinositol 3-kinase; PPARγ, peroxisome proliferator-activated receptor-γ; PTB, phosphotyrosine binding domain; ROS, reactive oxygen species; RXR, retinoid X receptor; SH2, src homology domain; TCA, tricarboxylic acid cycle; TF, transcription factor; CPT1, carnitine palmitoyltransferase-1; PTDETN, phosphatidylethanolamine.

Another reporter metabolite identified in our analysis is diacylglycerol (DAG), a lipid signaling molecule known to inversely correlate with insulin sensitivity (Holland et al, 2007; Itani et al, 2002; Roden, 2005; Savage et al, 2007; Shulman, 2000). Our results suggest that perturbations in DAG levels may be accompanied by changes in the adjacent CDP-Choline branch of the Kennedy pathway of phospholipid metabolism (**Figure 5.4**). Thus, DAG could potentially affect insulin sensitivity *via* activation of serine/threonine kinases or alterations in phospholipid membrane composition, both of which could lead to defects in insulin signaling, reduced insulin-stimulated glucose uptake, and glycogen synthesis- key metabolic features of diabetes (Muoio & Newgard, 2008) (**Figure 5.4**). Together, identification of these lipid-linked regulatory motifs and reporter metabolites known to be involved in type 2 diabetes pathogenesis provides further support for the validity of our approach.

Central carbon metabolism

Using our approach we found several reporter metabolites from the TCA cycle (citrate, 2-oxoglutarate, succinyl-CoA, fumarate and malate) (**Figure 5.4**). The down-regulated genes associated with these metabolites support the idea that TCA cycle and/or oxidative phosphorylation flux is reduced in diabetes (Patti et al, 2003). It is also interesting that ATP is one of the reporter metabolites, as the majority of cellular ATP is generated *via* respiration. Moreover, significant enrichment of binding motif for NF-κβ in the upregulated ATP neighbors is consistent with the potential role of this transcription factor in mediating oxidative stress responses triggered by by-products of incomplete β-oxidation (Sen & Packer, 1996). Another interesting finding is the enrichment of CREB family and NRF-1 motifs in enzymes associated with ATP and ADP. These results corroborate the role of CREB as an indirect regulator of nuclear-encoded oxidative phosphorylation genes *via* PGC1-α and other regulators linked to nuclear-encoded mitochondrial genes (**Figure 5.4**) (Patti et al, 2003; Scarpulla, 2006; Scarpulla, 2008).

The appearance of highly connected metabolites, such as ATP and NADH, among top-ranking reporter metabolites provides a possible link to the observed network-wide transcriptional changes in IGT and T2DM. Cellular levels of these co-factors are usually constrained within relatively narrow ranges to maintain thermodynamic stability. Oxidative phsophorylation, which is connected to TCA cycle flux *via* succinate and fumarate, accounts for most of the ATP (and NADH) turnover in a respiring cell. Our results suggest reduction in the activity of both TCA cycle and oxidative phosphorylation, in agreement with recent NMR data demonstrating that mitochondrial ATP synthesis is reduced in humans with insulin resistance (Petersen et al, 2003; Petersen et al, 2005; Szendroedi et al, 2007). Another major source of ATP and NADH production in the cell is glycolysis.

Reporter metabolites representative of glycolysis (glucose, glucose-6-phosphate, glucose-1-phosphate and pyruvate) also exhibited concordant down-regulation of the neighboring genes.

The concordance between the changes in gene expression levels for glycolysis, TCA cycle and oxidative phosphorylation in IGT and T2DM suggests that transcriptional regulatory mechanisms may be a response to altered levels of ATP/NADH. Such response may achieve two purposes: i) regulation of metabolism on global scale, as these co-factors are critical components of many metabolic pathways, and ii) regulation of NADH levels may help in reducing excessive (and potentially deleterious) oxidative stress resulting from sustained oxidation of excessive nutrients (Ristow et al, 2009). Although the way such regulatory control is mechanistically linked to the corresponding metabolites cannot be deduced from the gene expression data alone, there are several examples where metabolite co-factors are directly involved in regulating gene expression, *e.g.* NADH(/+) dependent regulation of genes in gram-positive bacteria (Brekasis & Paget, 2003), yeast (Anderson et al, 2003; Lin et al, 2000; Zhang et al, 2002) and human (Agarwal & Auchus, 2005; Rutter et al, 2001). NAD+ dependent changes in gene expression levels could also be mediated by the action of PGC-1α and SIRT1 complex, which have important roles in regulation of glucose homeostasis (Rodgers et al, 2005). Additional regulatory links, between glycolytic flux, energy metabolism, TCA cycle flux and fatty acid metabolism are also known in other eukaryotic systems such as baker's yeast (Cimini et al, 2009; Raghevendran et al, 2006; Schuurmans et al, 2008). Furthermore, several of the enzymes from central carbon metabolism may be regulated to a large extent at the post-transcriptional level (Daran-Lapujade et al, 2007; He et al, 2001). Parallels of such regulatory circuits in human cells may be discovered in the future with the here-identified transcription factors (Table S7) as one of the starting points.

Other pathways

Metabolites involved in protein and lipid glycosylation were found as reporters and characterized by down-regulation of neighboring enzymes (Table S2). Alterations in glycosylation may ultimately cause misfolding of several proteins, a feature previously associated with over-nutrition in hepatocytes (Shulman, 2000). Another reporter metabolite, shared by T2DM *vs.* NGT and T2DM *vs.* FH-comparison, is trichloroethanol, a metabolite in the cytochrome P450-mediated pathway derived from trichlorethene (Bruning et al, 1998). Although tricholoethanol or tricholoethene is not an endogenous metabolite in human tissues, it appears that the expression of the cytochrome P450 is altered in T2DM. Interestingly, experimental evidence shows that mouse exposure to trichlorethene leads to PPARα activation and the reprogramming of gene expression, resulting in induction of

enzymes mediating β- and ω-oxidation of fatty acids, and increased expression of genes involved in lipid metabolism (Laughter et al, 2004), a pattern similar to the T2DM metabolic phenotype (Shulman, 2000).

Reporter metabolites and macroscopic physiological parameters

The identification of reporter metabolites from glycolysis and energy-generation pathways suggests that there may be regulation of certain physiological parameters, such as glucose uptake, at the transcriptional level of the corresponding metabolic pathways. To investigate the extent of such possible regulation, we calculated Pearson correlation coefficients between insulin sensitivity (as measured by either whole-body glucose uptake during the hyperinsulinemic euglycemic clamp or insulin levels achieved during the OGTT) and mean centroid expression levels of genes surrounding reporter metabolites (Swedish dataset) (Materials and methods). A significant linear correlation with whole-body glucose uptake was observed for several reporter metabolites. In most cases, the correlation was significant only for one of the conditions (NGT, IGT or T2DM). For example, significant correlation of transcriptional regulation around dUDP with glucose uptake was found only for NGT samples (**Figure 5.5**A). It appears that this potential connection is de-linked under IGT and T2DM conditions. Another example is 1-Phosphatidyl-1D-myo-inositol 3-phosphate (**Figure 5.5**B), where significant correlation is observed with insulin level only for IGT. Further investigation of the causal mechanisms behind these observed correlation patterns may help in elucidating the regulatory role of the reporter metabolites in diabetes pathogenesis.

Potential biomarkers and pharmacological targets

A key scientific and clinical challenge is to identify molecular markers of diabetes risk, not only to better understand disease pathophysiology, but also to develop novel therapies for prevention and treatment of established diabetes. In this context, it is interesting that our analysis identified both PPARγ and its potential lipid ligands as regulatory molecules, since PPARγ ligand thiazolidinediones are currently employed as effective therapy for diabetes. We hypothesize that some transcriptional pathways identified in the current analysis, including CREB, NRF-1 and SRF, may be additional novel molecular mediators of the transcriptomic phenotype associated with insulin resistance, and thus potential targets for future intervention strategies. Of course, the potential roles of these pathways will require additional testing in cultured cells and animal models, where their impact on metabolic flux and insulin sensitivity can be fully assessed.

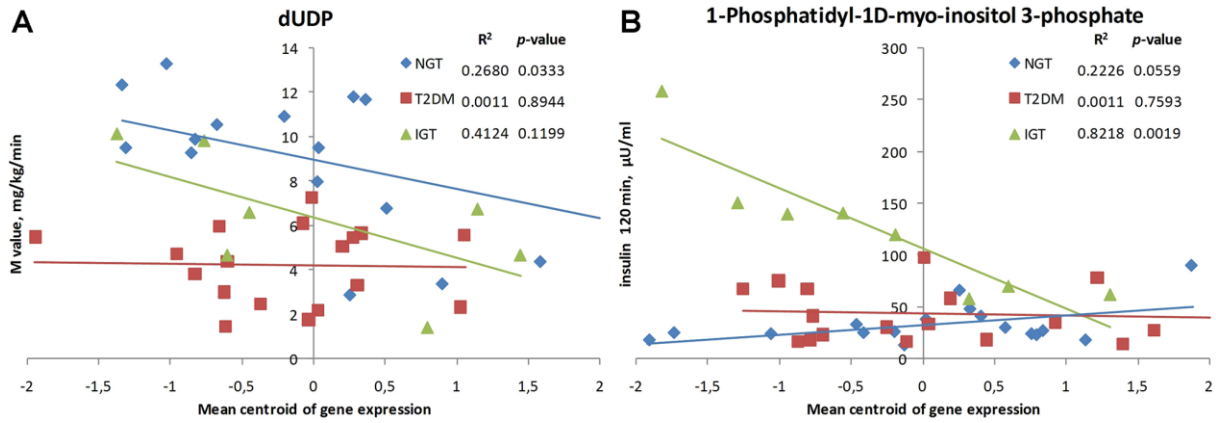


Figure 5.5. Correlation of glucose uptake and insulin level with mean centroid expression levels of reporter metabolite neighbor genes (Swedish male dataset). M value – whole-body glucose uptake during the hyperinsulinemic euglycemic clamp, Insulin 120 min – insulin levels achieved at the two hour time point of oral glucose tolerance test.

Similarly, reporter metabolites identified in our analysis represent molecules likely to be involved in human skeletal muscle insulin resistance phenoytpes and also novel candidate biomarkers of insulin resistance and diabetes risk. In support of this hypothesis, several of the identified metabolites have known physiological roles in T2DM (Table S8 and Discussion above). Additional molecules have been analyzed either in rodents and/or in other tissues (Table S8) and thus, their appearance as reporter metabolites also strongly implicates their involvement in insulin resistance in human skeletal muscle. Some of the more novel metabolites identified in our analysis, including glycolytic and fatty acid oxidation intermediates, are known targets of metformin, a compound effective for diabetes therapy and prevention (**Figure 5.4**). We also identified an interesting link between DAG, a reporter metabolite for T2DM, and the CDP-Choline branch of the Kennedy pathway of phospholipid metabolism (**Figure 5.4**). This pathway has been implicated in cancer development and is being established as anti-tumor drug target (Banez-Coronel et al, 2008; Ramirez de Molina et al, 2005). Changes in phospholipid metabolism are also known to affect the properties of cellular membranes, and subsequently signaling through membrane proteins. Further investigation of the role of phospholipids in T2DM pathogenesis may provide clues to some of the missing links that connect metabolic flux changes with insulin signaling in skeletal muscle cells.

Supplementary tables S1-S4 list additional reporter metabolites which are, to our knowledge, not (directly) linked with any of the known metabolic players in T2DM. Our analysis nevertheless suggests them as potential nodes of disruption or as biomarkers. Measurement of the intramyocellular concentration of the reporter metabolites in patients with diabetes risk may help to confirm the role of these metabolites in insulin resistance.

Metabolic hubs as reporters

A particularly interesting finding from our analysis is the identification of highly connected metabolites as reporters, including ATP/ADP and NAD+/NADH. We hypothesize that diverse environmental and genetic risk factors result in insulin resistance when individuals are unable to mediate appropriate compensatory transcriptional and metabolic responses in other parts of the network connected by these hubs. Our results also suggest that alterations in gene expression linked to the highly connected co-factors are likely to be acquired features of established T2DM. Analysis of the transcriptional activity of CREB in the context of ATP concentrations and TCA cycle activity in skeletal muscle may help to elucidate regulatory mechanisms leading to these changes.

Constraints and extension of methodology

Reconstructed human metabolic network models are still evolving, incomplete, and subject to error. Well-annotated pathways such as central carbon metabolism are thereby likely to be over-represented in the reporter analysis. In order to partially compensate for this limitation, we used two reconstructions – Recon1 and EHMN. As network reconstructions will become more complete, it will be possible to better assess the extent of this limitation. Another essential input to our algorithm, in addition to metabolic network, is gene expression data for the genes represented in the network. We would like to note that neither EHMN nor Recon1 network genes were fully represented by the microarray chips used in the two case studies. Only 54% and 39% genes from the Recon1 and EHMN, respectively, were represented on the chips used in Mexican-American case study, while this coverage was 85% and 60% in Swedish case study. Interestingly, re-analysis of the Swedish Male dataset by using only a subset of genes from the HG-U133A chip that were represented also on the HuGeneFL (used in Mexican-American case study) showed a large overlap between the two reporter metabolite sets thus obtained (86% for T2DM *vs.* NGT comparison and 69% for the rest two). The details of this analysis, together with relevant statistical considerations, can be found in **Supplementary Figure 5.4, Supplementary Table 5.1**, **Supplementary Table 5.2**.

Although the present analysis identified common metabolic and regulatory signatures across the two studies, there are several differences in the study designs, and therefore the results must be regarded with certain caution. In addition to relatively low number of subjects in Mexican-American study, the differences include fasting state biopsies in Mexican-American study *vs.* post insulin stimulation biopsies in Swedish study. Furthermore, the age and BMI (Body Mass Index) of the individuals participating in the two studies were different and may contribute to the differences in the observed gene expression patterns. To our knowledge, these two case studies represent the only

human skeletal muscle transcriptome datasets that were publically available at the time of here reported computational analysis. Analysis of new datasets which may become available in the future will be useful in obtaining further insight into molecular physiology of skeletal muscle in the context of T2DM. Moreover, emergence of better or new gene expression analysis tools will help to cover parts of metabolic network that are currently inaccessible due to the lack of data.

Extension of the analysis to discover more global regulatory patterns by using additional bio-molecular interaction data (Oliveira et al, 2008) such as protein-DNA and protein-protein interactions will definitely be an important step in obtaining a higher resolution picture of T2DM metabolic phenotypes. Availability of such interaction data at the high confidence level of metabolic interactions is the current major bottleneck. Another essential extension of the methodology will require the use of thermodynamic data for metabolic reactions (Cakir et al, 2006; Henry et al, 2007; Kummel et al, 2006). Moreover, since mRNA levels do not necessarily correlate with the protein levels, incorporation of the proteomics data together with the thermodynamic data will allow more accurate interpretation of the reporter metabolites in terms of implications for flux and concentration changes.

Conclusions

We demonstrate the use of a network-guided data integration approach to discover key, physiologically relevant metabolic and regulatory nodes in T2DM pathogenesis. The methodology does not require the use of *a priori* disease-specific knowledge regarding the involvement of specific pathways or metabolites, thereby making it a robust and unbiased analytical framework for studying diseases linked to perturbations in the cellular metabolic network. Our results identify the highly connected metabolites ATP and NAD as reporters and potential mediators of the widespread changes in gene expression linked to insulin resistance in muscle. Moreover, our results extend previous knowledge about T2DM pathogenesis at the gene expression level- by reporting additional potential sites of disruption, *e.g.*, TCA cycle and Kennedy pathway of phospholipid metabolism. Several metabolites from other pathways were also found to display significant differential gene expression of the genes around them and we suggest putative regulatory mechanisms behind these alterations. Our results suggest a framework of metabolic disruption observed with insulin resistance and diabetes, which can be used to test the role of specific pathways in mediating disease pathophysiology, and more practically, for the identification of potential biomarkers for preventive and therapeutic monitoring.

Materials and Methods

Gene expression and sequence data

Two datasets used in the study were obtained from the Diabetes Genome Anatomy Project website (http://www.diabetesgenome.org). Brief comparison of microarray platforms from the experimental studies (Mootha et al, 2003; Patti et al, 2003) used in the current work is presented in the **Supporting Text 1**. Promoter sequences for all genes were obtained from the Ensembl Biomart (http://www.ensembl.org/biomart). The transcriptional start sites (TSSs) were identified based on the annotation of the Ensembl Biomart sequences. Sequences in the -800 to 200 base pairs region of the TSS were deemed as promoter regions for the analysis. Interspersed repeats and low complexity DNA sequences were masked out.

Metabolic networks

Two reconstructions of human metabolic network, *viz.*, Recon1 (Duarte et al, 2007) and EHMN (Ma et al, 2007) were used in this study. The *Homo sapiens* Recon1 is a comprehensive literature-based metabolic network reconstruction that accounts for the functions of 1496 ORFs, 2004 proteins, 2766 metabolites and 3311 metabolic and transport reactions. The ENMN (Edinburgh Human Metabolic Model) is a network reconstructed by integrating genome annotation information from different databases and metabolic reaction information from the literature. The network contains nearly 3000 metabolic reactions, which were reorganized into about 70 human-specific pathways according to their functional relationships. The two models mainly differ in the coverage of reactions and in the accounting of compartmentalization and inter-organelle transport reactions.

Significance of differential gene expression

Preprocessing of the gene expression data was carried out by using the statistical software environment – R (www.r-project.org). The probe intensities were obtained and corrected for background by using robust multi-array average method (RMA) (Irizarry et al, 2003 2003) with only perfect-match (PM) probes. Normalization was performed by using the quantiles algorithm. Gene expression values were calculated from the PM probes with the median polish summarization method (Irizarry et al, 2003 2003). All data preprocessing methods were used by invoking them through the *affy* package (Gautier et al, 2004) by using *rma* function. Significance of the differential expression was calculated by using the empirical Bayes test (Smyth, 2004). The probe-sets were grouped into genes, and to each gene the differential expression was defined by choosing the value from the top level probe-set (using the probe-set rank defined by Affymetrix). In case of more than one probe-set present at the top level, the median value was used.

Reporter metabolites

Each metabolite in the metabolic network was scored based on the scores of its *k* neighbor enzymes (*i.e.* enzymes catalyzing reactions involving that metabolite, either as a substrate or as a product). Each enzyme was assigned with a *P*-value for differential expression based on the *P*-value of the gene encoding for that enzyme. In case of isozymes and enzyme-complexes, genes with most significant expression change were used to score the enzyme (**Figure 5.1**). *P*-values of genes p_i, indicating the significance of differential expression, were converted to Z-scores Z_i by using the inverse normal cumulative distribution function (CDF) (θ^{-1}): $Z_i = \theta^{-1}(1 - p_i)$. All metabolite nodes were assigned a Z-score, $Z_{metabolite}$, calculated as aggregated Z scores of the *k* neighbor enzymes: $Z_{metabolite} = \frac{1}{k}\sum Z_{ni}$. $Z_{metabolite}$ scores were then corrected for the background distribution by subtracting the mean (μ_k) and dividing by the standard deviation (σ_k) of the aggregated Z scores derived by sampling 10000 sets of *k* enzymes from the network: $Z_{metabolite}^{corrected} = \frac{(Z_{metabolite} - \mu_k)}{\sigma_k}$.

Corrected Z-scores were then transformed to *P*-values by using CDF. Metabolites with *P*-values less than 0.05 were deemed as reporter metabolites. Detailed information on the reporter scoring can be found in the **Supporting Text 1** and (Patil & Nielsen, 2005).

Transcription factor binding site enrichment

For all reporter metabolites, we assessed enrichment of known protein-binding sequence motifs in the promoter regions (–800 to 200 base pairs relative to the transcription start site) of the corresponding neighbor genes. In order to obtain robust results, we only considered sets consisting of at least 5 up- or down-regulated genes. For each reporter metabolite, the sequences of all enzyme neighbors were used as the positive sequence set, whereas all other enzymes in the network model were used as the negative (background) set. Known motifs were identified by using position frequency matrices of all known motifs stored in the TRANSFAC database (Matys et al, 2003). The motif enrichment analysis tool ASAP (Marstrand et al, 2008) was used to scan all TRANSFAC motif matrices against the positive sequence sets of each reporter metabolite. The negative sequence sets were used together with 2nd order background model. A one-tailed Fisher's exact test was used to assess per-sequence over-representation of any known motif, and the threshold used to calculate significance for each TRANSFAC matrix was set to 70% of the highest-scoring sequence motif. The q-value cut-off criteria (Storey & Tibshirani, 2003) was used as a post-data measure of statistical significance of all motifs found to be significantly enriched.

Acknowledgements

We are thankful to Isabel Rocha for feedback on the manuscript text. We thank Anders Krogh for advising on the choice of motif analysis method. We are grateful to the reviewers for useful comments.

Supplementary information

Supporting Text 1 Calculation of reporter metabolite score

List of Supplementary Figures

Supplementary Figure 5.1 Conversion of gene *P*-values to standard Z-score using inverse cumulative function.

Supplementary Figure 5.2 Scoring system based on distribution of means.

Supplementary Figure 5.3 Scoring system for identification of reporter metabolites.

Supplementary Figure 5.4 Comparison of results from reporter metabolite analysis based on two different sets of metabolic genes

List of Supplementary Tables

Supplementary Table 5.1 Brief comparison of the two experimental studies used for identifying metabolic and regulatory signatures of T2DM.

Supplementary Table 5.2 Brief comparison of microarray platforms from experimental studies used for identifying metabolic and regulatory signatures of T2DM.

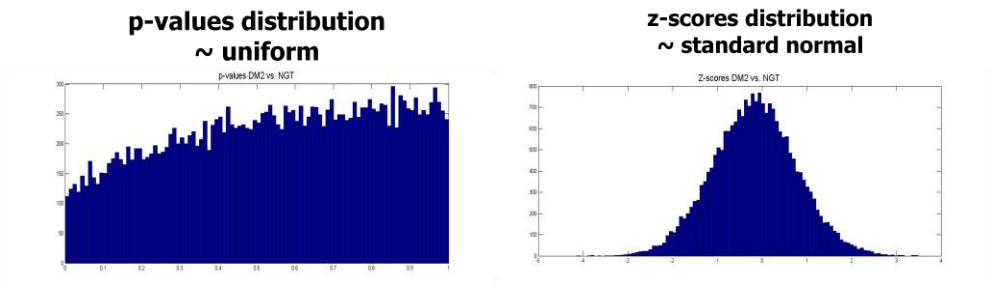
Tables S1-S8, including Figure S1 can be accessed through the web:

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000729#s5

Supporting Text 1

From gene expression to reporter metabolites – calculation of reporter metabolite score

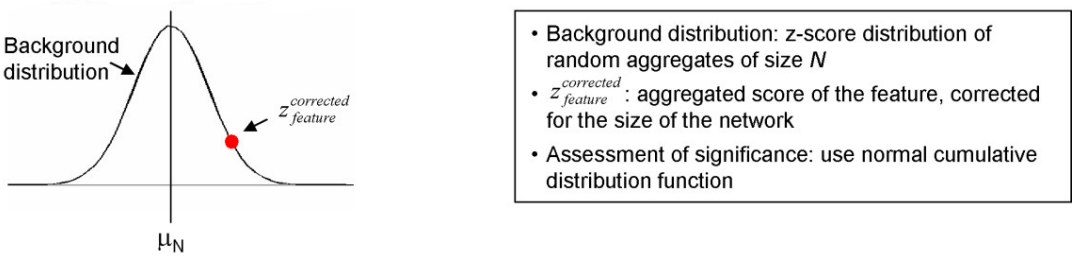
We present a step-by-step procedure for calculation of reporter metabolites. The first step is to estimate *P*-value for the significance of differential expression for all of the genes across two phenotypes in question, for example, T2DM *vs.* NGT. In our analysis we used empirical Bayes test (Smyth, 2004) to assess the significance of differentially expressed genes. It is possible to use other statistical methods, *e.g.* t-test or ranksum test. The choice of the method depends on the underlying structure of the data and assumptions that can be made. The statistical test assigns a *P*-value to each of the probe sets by taking into account the variation within the groups being compared. A gene is usually represented by several probe sets, and in such case a *P*-value for differentially expressed gene is defined by choosing the value from the top probe-set[1], according to the probe-set ranks (probe-set ranks are defined by the manufacturer, see Affymetrix web site- http://www.affymetrix.com/). Further, by using inverse normal cumulative distribution function, the *P*-value of each of the probe set (gene) is converted to a standard Z-score with a mean of 0 and a variance of 1, thus uniformly distributed *P*-values are transformed to a normally distributed random variable (**Supplementary Figure 5.1**).



Supplementary Figure 5.1 Conversion of gene P-values to standard Z-score using inverse cumulative function.

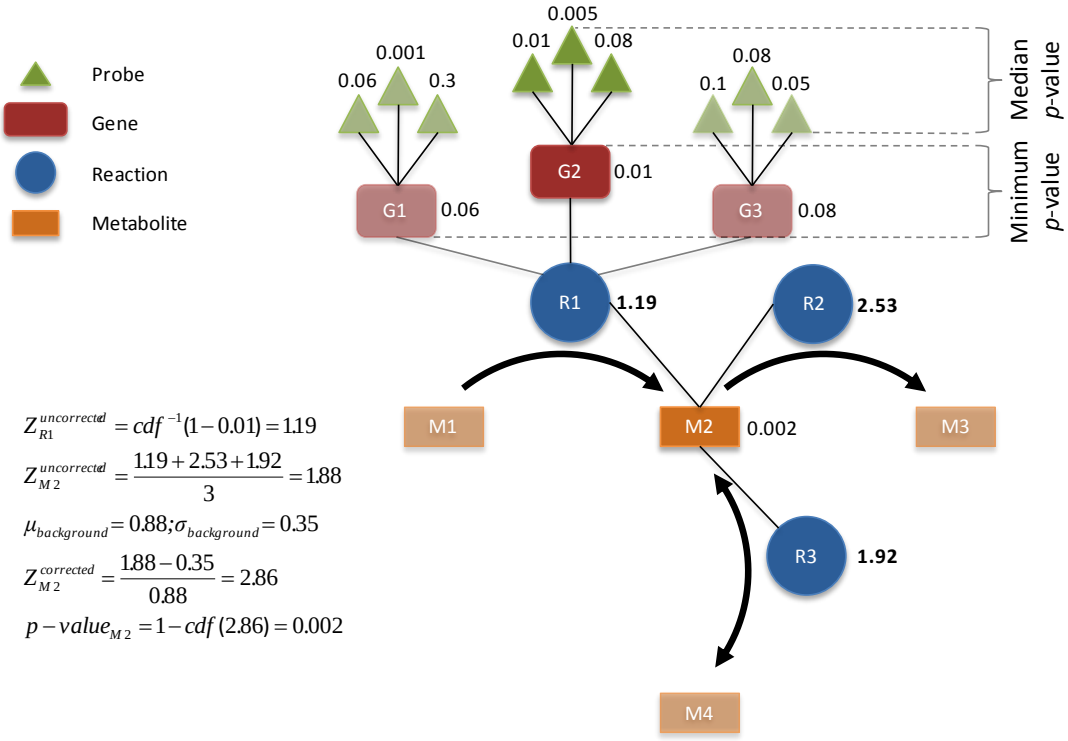
The next step is to use gene Z-scores to assign Z-score to the metabolites. Z-score of a metabolite is an average of Z-scores of genes connected to that metabolite in the metabolic model. In order to evaluate the significance of metabolite scores, average Z-score of each of the metabolites needs to be corrected for the background Z-score distribution[2] (**Supplementary Figure 5.2**).

[1] In case if there are several probe-sets representing the same rank the median *P*-value is selected
[2] See main text, materials and methods section for background distribution estimation.

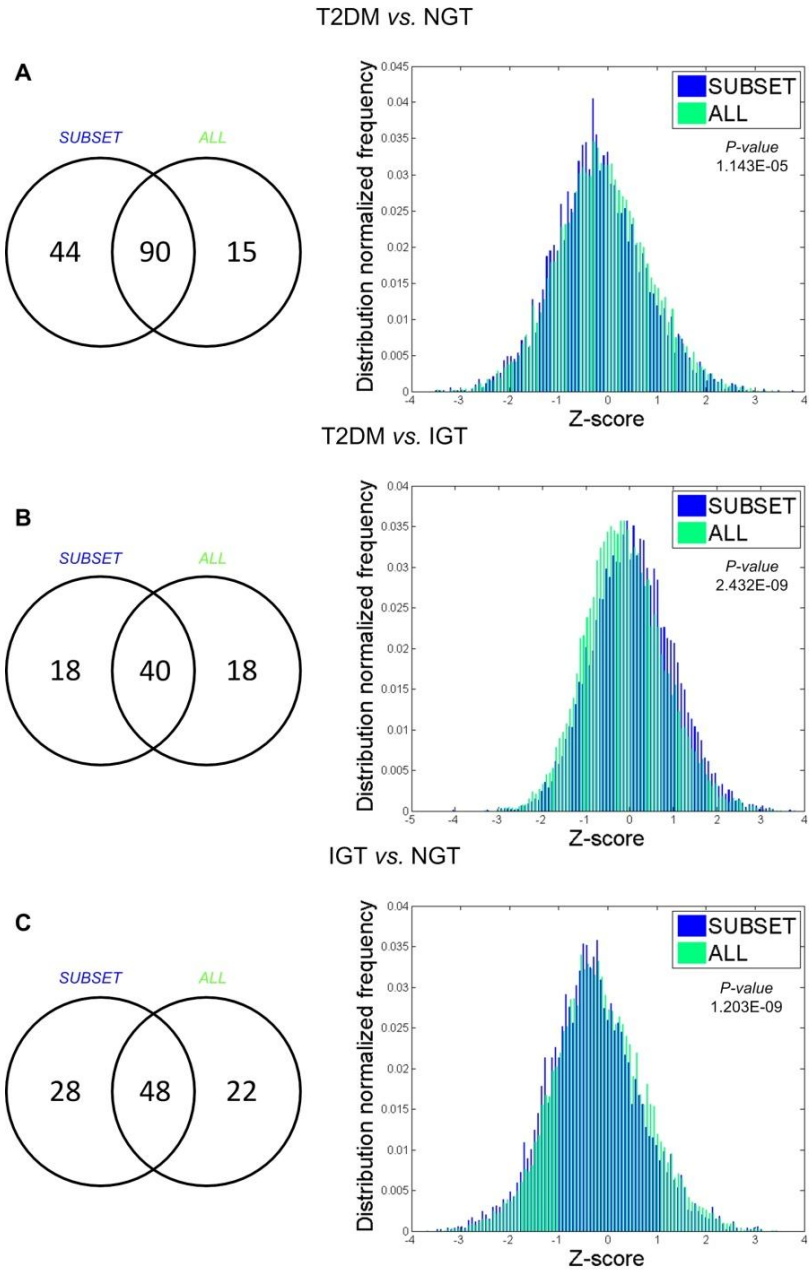


Supplementary Figure 5.2 Scoring system based on distribution of means. Here a feature is a metabolite. Illustration source: (Oliveira et al, 2008).

After background correction each metabolite's Z-score is converted to a *P*-value by using normal cumulative distribution function. Overall procedure including a calculation example is illustrated in **Supplementary Figure 5.3**. Metabolites having significant *P*-values are termed reporter metabolites – metabolites around which the most significant transcriptional changes occur. For readers interested in further details on the scoring system, including alternative scoring schemes and background distribution, please refer to (Oliveira et al, 2008).



Supplementary Figure 5.3 Scoring system for identification of reporter metabolites. Each metabolite is scored based on the scores of the associated enzyme-catalyzed reactions. Each enzyme, in turn, is assigned a score based on median of the *P*-values of the probes representing the corresponding gene. In case of a reaction catalyzed by an enzyme complex or a set of isozymes, minimum of the *P*-values of the corresponding enzymes is chosen. Numbers in bold are Z-scores for each reaction, the rest of the numbers represent *P*-values (significance of differential expression).



Supplementary Figure 5.4 Comparison of results from reporter metabolite analysis based on two different sets of metabolic genes – one being proper subset of the other. In order to assess the effect of relatively low coverage of metabolic genes on the HuGeneFL (used in Mexican-American case study), we re-analyzed the Swedish Male dataset by using only a subset of genes from the HG-U133A chip that were represented also on the HuGeneFL. The results show large overlap between the two reporter metabolite sets thus obtained. The left panel shows Venn diagram illustrating the overlap between the two reporter metabolite sets, while the right hand side panel shows the distribution of Z-scores for the two metabolic gene sets. *P*-values shown are results of Student's t-test comparing the two distributions. In all comparisons, *P*-values are very low, implying that the two distributions are distinct. This is one of the contributing factors to the difference between the reporter metabolite results, in addition to the fact that the number of neighbors for several metabolites is also different for each gene subsets. For mathematical description of the relationship between these two factors and reporter score, please see materials and methods section in the main text.

Supplementary Table 5.1 Brief comparison of the two experimental studies used for identifying metabolic and regulatory signatures of T2DM.

	Swedish male dataset (Mootha et al, 2003)			Mexican-American dataset (Patti et al, 2003)		
Phenotype	**NGT**	**IGT**	**T2DM**	**FH-**	**FH+**	**T2DM**
Ethnicity	Caucasian			Mexican American		
Subject Number	17	8	18	6	4	5
Age	66.1 (3.4)	66.4 (1.6)	65.5 (1.8)	38.5 (3.7)	40.8 (2.6)	43.8 (2.1)
BMI, kg/m^2	23.6 (3.4)	27.1 (4.8)	27.3 (4.0)	31.2 (0.8)	28.9 (1.4)	37.4 (5.8)
Microarray	Affymetrix HG-U133A			Affymetrix HuGeneFL		

Supplementary Table 5.2 Brief comparison of microarray platforms from experimental studies used for identifying metabolic and regulatory signatures of T2DM.

Microarray platform	HG – U133A	HuGeneFL (Hu6800)
Number of genes	>14 500	~6800
Number of probe sets	>22 000	6940
Recon1 coverage*, %	85%	54%
EHMN coverage*, %	60%	39%

*Percentage of genes from the model represented on the microarray chip.

"In the animal world we have seen that the vast majority of species live in societies, and that they find in association the best arms for the struggle for life: understood, of course, in its wide Darwinian sense – not as a struggle for the sheer means of existence, but as a struggle against all natural conditions unfavorable to the species. The animal species, in which individual struggle has been reduced to its narrowest limits, and the practice of mutual aid has attained the greatest development, are invariably the most numerous, the most prosperous, and the most open to further progress. The mutual protection which is obtained in this case, the possibility of attaining old age and of accumulating experience, the higher intellectual development, and the further growth of sociable habits, secure the maintenance of the species, its extension, and its further progressive evolution. The unsociable species, on the contrary, are doomed to decay."

— Peter Kropotkin, Mutual Aid: A Factor of Evolution, 1902

Chapter 6 Co-occurring bacterial communities feature high potential for metabolic cooperation

Abstract*

Microorganisms in their natural habitat often live within large communities. Properties of the individual species within a microbial community as well as interspecies interactions are fundamental in determining its collective function and ecological impact. While microbial communities from diverse habitats are increasingly being characterized in terms of the constituent member species, our knowledge of the interspecies molecular interactions remains largely incomplete. Towards addressing this question, we identified 7221 microbial groups, with up to 4 members, which co-occur across 1237 metagenomic samples. We then used a multi-species network modeling approach to estimate metabolic interaction potential and interspecies dependencies within microbial communities. Here we show that microbial communities that co-occur across diverse environments have markedly higher potential for metabolic interactions. The co-occurring communities showed a higher number of possible mutualistic metabolic exchanges and higher degree of interspecies dependency. In comparison, phylogenetic relatedness of the member species and their similarity in nutritional requirements were both found to be poor descriptors of co-occurrence, suggesting that the potential for cooperation outweighs the risk of resource competition in determining community structure. The concept of metabolic interaction potential introduced here has implications for understanding and manipulation of the structure and stability of natural as well as synthetic microbial communities.

* Manuscript in preparation: Zelezniak A, Andrejev S, Ponomarova O, Patil KR. Co-occurring bacterial communities feature high potential for metabolic cooperation

While molecular biology of monocultures is studied at a high level of detail, interspecies interactions and ecological dependencies that exist in mixed microbial populations remain poorly understood. Microbial communities of ecological and medical importance are currently probed mainly for the identification of the composing members. The knowledge of interspecies interactions, however, is lacking in the case of microbial communities living in their natural habitats, as these are difficult to investigate *in situ*. It is generally believed that nutrient exchanges make up a considerable fraction of interspecies interactions in natural microbial communities (Cox et al, 1974; Moller et al, 1998; Phelan et al, 2012). Interspecies metabolic interactions can take various forms, such as synthrophy (sequential metabolism of degradation products) (Falony et al, 2006; Ze et al, 2012), mutualism (Periasamy & Kolenbrander, 2009), commensalism (Rakoff-Nahoum et al, 2004) or resource partitioning (Lawrence et al, 2012). Such metabolic exchanges can confer several advantages to the community as a whole, for example, more efficient and complete use of available nutrients (Poltak & Cooper, 2011), or increased ability to survive under diverse/changing nutrition availability, or higher resistance to stressors (Ramsey et al, 2011). The current knowledge on metabolic interactions in microbial communities, however, is limited to synthetic communities with two (occasionally three) members (Freilich et al, 2011; Miller et al, 2010; Stolyar et al, 2007; Wintermute & Silver, 2010). Thus, the question remains open regarding the contribution of metabolic interactions in shaping the architecture of natural communities, which consist of from tens to hundreds of species. We address this question in two steps. First, we take advantage of the large number of metagenomics studies available to estimate groups of microbial species that significantly co-occur in different habitats/samples. Co-occurrence of such a microbial community suggests functional relationships within the community members. In the second step, we developed a multi-species metabolic modeling approach and thereby tested the hypothesis that the observed co-occurrence patterns are largely due to interspecies metabolic interactions.

Darwin originally recognized that phenotypic similarity of species influence their interaction with other species and environment in predictable ways (Cavender-Bares et al, 2009). In particular, Darwin noted that if closely related species are evolutionary similar, they should share similar environmental requirements and thus would be expected to co-occur. On the other hand, closely related species should exhibit strong competitive interactions due to their ecological similarity, thereby limiting their coexistence. Thus, in present work, we wanted to investigate this paradoxical phenomenon on the example of microbial communities. Specifically, we were interested in examining the roles of competition and cooperation in co-occurrence of microbial species by

comparing their phylogenetic relatedness and their metabolic exchange capacity under austere nutritional conditions.

Results

Determining co-occurring lineages

Microbial co-occurrence studies so far have been restricted to two species at a time. Such binary associations, although insightful (Chaffron et al, 2010; Freilich et al, 2011), may not fully reflect the underlying functional interactions/dependencies, as the latter can be masked by the interactions mediated by other species. Consequently, such indirect interactions can hamper the interrogation for functional molecular interactions, *e.g.* metabolite exchange. We therefore started by compiling a list of co-occurring communities consisting of up to four species by using Fischer's exact test, applied to a dataset consisting of 2801 metagenomic samples (Methods). While 16S rRNA groupings into operational taxonomic units (OTU) provide a reasonably good estimate of phylogenetic diversity among samples, they do not carry any information about molecular functions of lineages. To address this issue and to directly investigate species co-occurrence patterns in environmental samples, we assigned OTU representative sequences (Methods) to its best match 16S rRNA gene from fully sequenced bacterial genome. In total, about 11% of all OTUs (OTU defined at 97% sequence identity, (Chaffron et al, 2010)) appearing at least 3 times among the sampling sites were mapped to 261 bacterial genomes given the stringent criteria used for mapping 16S rRNA sequences to the bacterial species (Methods). 7221 significantly co-occurring communities (FDR ≤ 0.01) were identified in this fashion: 381 pairs, 3322 triplets and 3518 of quadruplets (four members) (**Supplementary Table 6.1**).

Co-occurring microbial lineages have similar nutritional requirements

The recent automated genome-scale metabolic reconstruction pipeline Model SEED (Henry et al, 2010) allowed us to reconstruct simulation-ready bacterial models for all of the mapped genomes (Methods). Although Model SEED provides simulation ready models, reconstructed metabolic networks, especially non-curated, are still evolving, incomplete, and potentially subject to error. To partially address these problems we made a few manual changes (Methods). For example, some reaction reaction's directions varied among different automatically reconstructed models. To overcome such an inconsistency, we used directions of the maximum possible number of reactions from 8 manually reconstructed models, which still kept model suitable for simulations (Methods).

Given that we observed combinations of lineages (communities), which were overrepresented in samples spanning different habitats, we hypothesized that co-occurring individuals would tend to

have similar nutritional requirements. To corroborate this, we developed a metric which shows for any given number of species an overlap of common growth requirements for each individual in a given community. In other words, metabolic resource overlap (MOP) represents a similarity of species with regards to their resource requirements to produce individual biomasses (Methods). We computed metabolic resource overlaps for co-occurring communities and compared them with random combinations of species. Indeed, we observed that within co-occurring pairs and quadruplets metabolic requirements were significantly more similar than the random communities (*P*-value = 0.004; *P*-value < 10^{-7} correspondingly, Wilcoxon rank-sum test), though the observed difference in median from the random control was not substantial (**Figure 6.1**A, **Supplementary Figure 6.1**A).

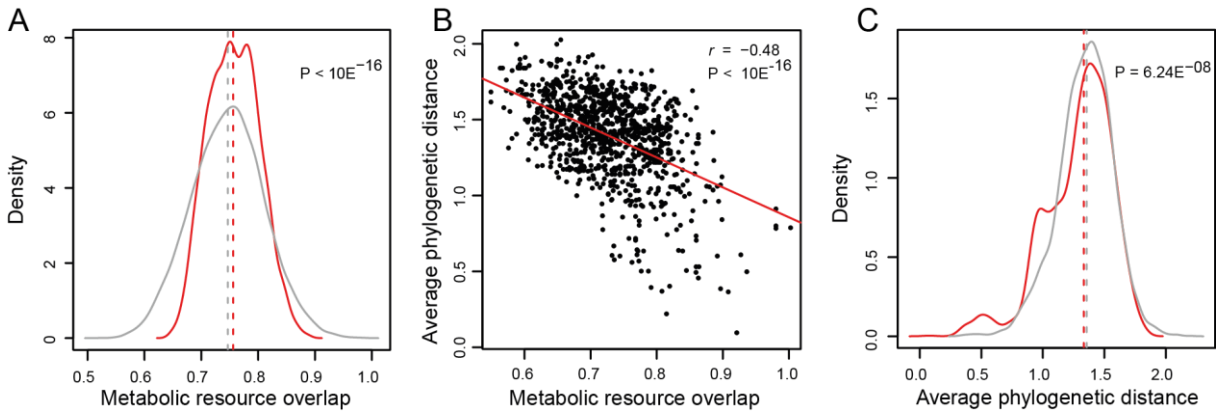


Figure 6.1 Species metabolic requirements are reflected in evolutionary divergence. A) Species with similar metabolic requirements tends to co-occur. Density plot represents distribution of calculated metabolic resource overlaps for co-occurring quadruplets (red curve), see **Supplementary Figure 6.1**A for all results. B) Phylogenetically close species have more similar metabolic requirements. Red line is least squares linear regression fit C) Average phylogenetic distance is significantly smaller in co-occurring communities (4 species communities presented, see **Supplementary Figure 6.1**B for all results). Control was computed by randomly sampling 10000 times combinations of groups consisting 4 members (grey curve). Importantly, random sampling was performed only from mapped genome space (261 metabolic reconstructions). Dotted lines represent medians of distributions.

One global, though indirect, metric of functional diversity of a microbial community is the phylogenetic diversity of its members. One would expect if two species are evolutionarily very similar, they would have had similar needs, and therefore their minimal growth requirements would be alike. Evolutionary distance is a complex biological phenomenon which is to a large extent is reflected in the DNA sequence diversity; however, it is not very obvious to what extent phylogenetic distance reflects the metabolic signal. We compared our estimation of shared metabolic resources for 10,000 groups of 4 species to the average phylogenetic distance between species in corresponding groups (**Figure 6.1**B). The observed significant (r = -0.48, *P*-value < 10^{-7}) correlation between phylogenetic distance and metabolic resource overlap metric supports the biological validity of metabolic models, and the fact that closely related species tend to have similar metabolic

requirements. Subsequently, as we did for metabolic resource overlap, we compared phylogenetic distance in co-occurring communities to those of the randomly grouped species. Indeed, we observed that phylogenetic distance in the co-occurring pair and quadruplet communities was significantly less than the random control (*P*-value = 0.00007; *P* -value < $6.24 \cdot 10^{-8}$ correspondingly, Wilcoxon rank-sum test, **Figure 6.1**C, **Supplementary Figure 6.1**B). The difference, however, was only moderate, and the three-species communities did not show significant difference (*P*-value = 0.783, Wilcoxon rank-sum test, **Supplementary Figure 6.1**B). Overall, neither metabolic resource overlaps nor phylogenetic distances were substantially different from random groups of species.

Co-occurring bacterial communities feature high potential for metabolite interactions

Metabolite cross-feeding provides an opportunity for species to compensate for the lack of nutrients in the environment (Wintermute & Silver, 2010), and thus can induce fitness of community members. In a group, individuals will have the maximum number of interactions under growth conditions which are favorable for the whole group but not for each individual. Such a condition would be found in minimal media (in terms of number of components presented) and not necessary unique one which supports growth of every individual in community (Klitgord & Segre, 2010). Therefore, we wanted to assess the impact of living in community from the growth perspective. If we imagine a community living without interactions it would require more nutrients to sustain each individual's growth as opposed to cross-feeding community. The difference in minimum number of components presented in medium between non-interacting and interacting community demonstrates the impact of living in a group (**Figure 6.2**A).We defined this difference, how many fewer components would the community need if it was interacting, as metabolic interaction potential (MIP) (**Figure 6.2**B, Methods).

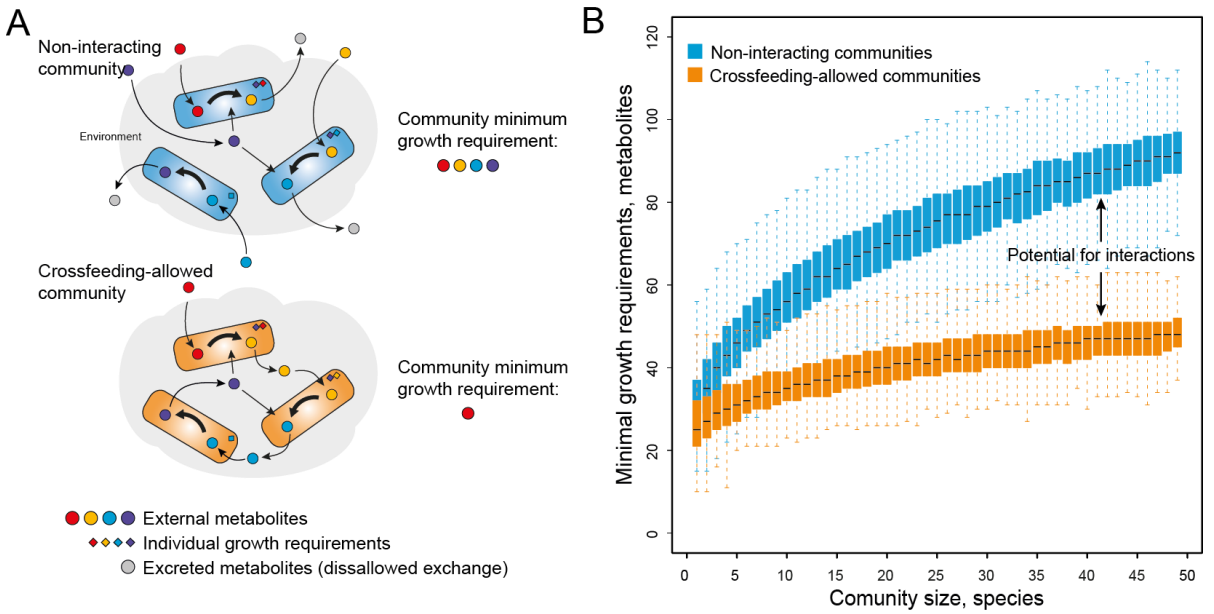


Figure 6.2 Interacting communities potentially require fewer components needed for growth. A) A community without interaction minimally needs 4 metabolites from environment to have all growing members (upper part). In contrast, a crossfeeding group of species potentially would require only one component to grow (lower part). B) Crossfeeding communities have lower minimal growth requirements potentially inducing interaction within community. Each boxplot represents a distribution of minimal growth requirements for random group (sampled 1000 times) of species of a given size. Random sampling procedure was performed using 1532 metabolic models space.

Although metabolic resource overlap (MOP) and phylogenetic distance in a few cases demonstrate significant trends, neither of them was spectacularly different from random control nor provided a clue about nature of interactions (**Supplementary Figure 6.1**). Thus, we wanted to investigate whether metabolic interaction potential could explain observed microbial co-occurrence patterns. In contrast to MOP and phylogenetic distance, for all the combinations of co-occurring groups of species the metabolic interaction potential showed significant strong signals (*P*-value = $2.58 \cdot 10^{-5}$; *P*-value < 10^{-7}; *P*-value < 10^{-7} respectively for pairs, triplets and quadruplets, Wilcoxon rank-sum test) (**Figure 6.3**A).

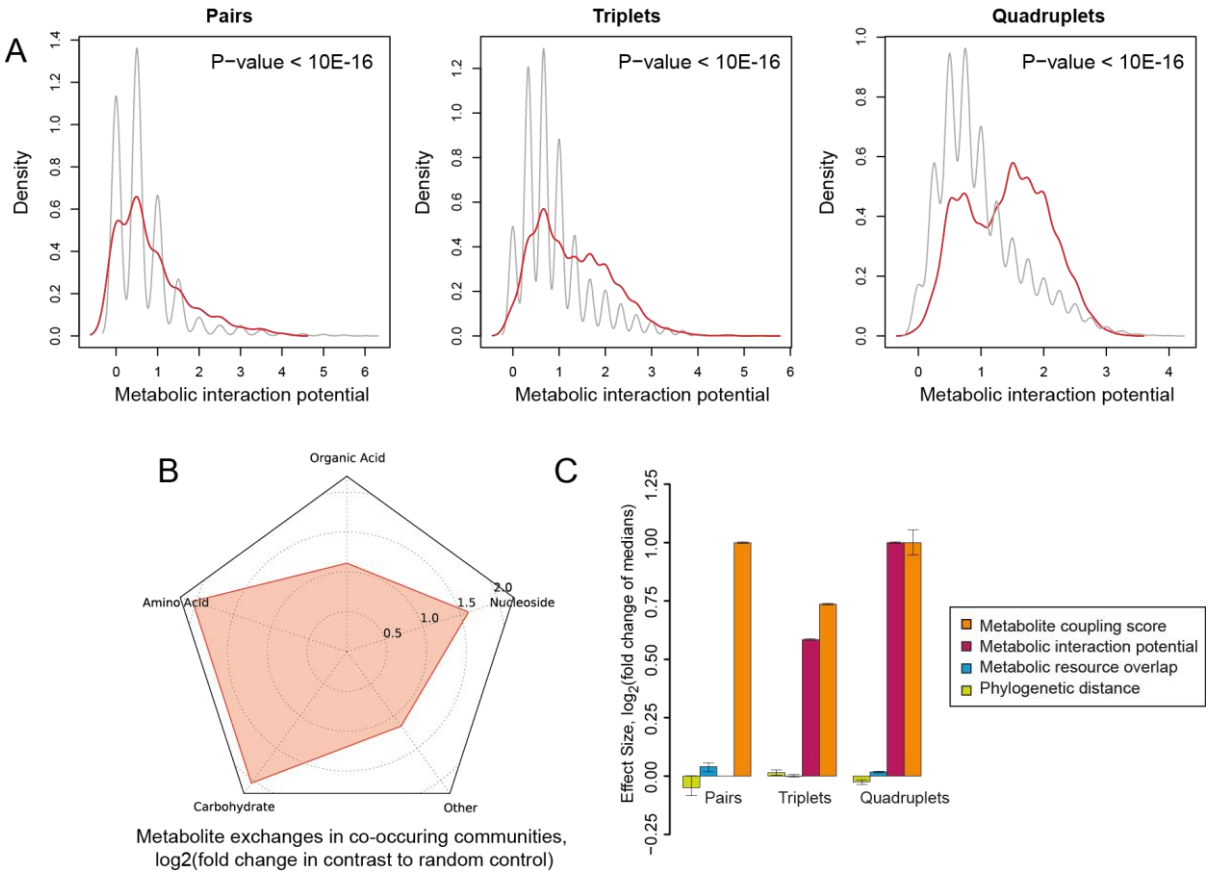


Figure 6.3 Potential for cooperation outweighs the risk of competition. A) Metabolic interaction potential is substantially higher within co-occurring communities (red curves), especially larger difference observed in groups of 4 members (right panel). Controls were computed by randomly sampling species combinations of corresponding group sizes 10000 times (grey curves). Importantly, random sampling was performed only from mapped genome space (261 metabolic reconstructions). B) Spider plot demonstrating the importance of metabolite exchanges in co-occurring communities (triplets example). The exchanges of amino acids and carbohydrates are needed more frequently in co-occurring communities in contrast to random control. C) Species metabolite coupling scores and metabolic interaction potential is higher in co-occurring groups, indicating the importance of cooperation for microbial coexistence.

Notably, a better signal resolution was observed if we considered higher combinations of species (combinations of four in contrast to triplets and pairs) suggesting that metabolic interactions can occur in higher order multi-species communities (**Figure 6.3**A, **Supplementary Figure 6.1**). To corroborate this, we performed simulations by computing metabolic interaction potential for randomly chosen species combinations of different sizes from space spanning 1503 bacterial metabolic reconstructions representing a majority of currently known bacterial taxa. We observed a decaying trend with a peak interaction potential at a community size of 6 species, supporting our finding that that metabolic interactions could possibly exist in higher order microbial communities (**Figure 6.4**).

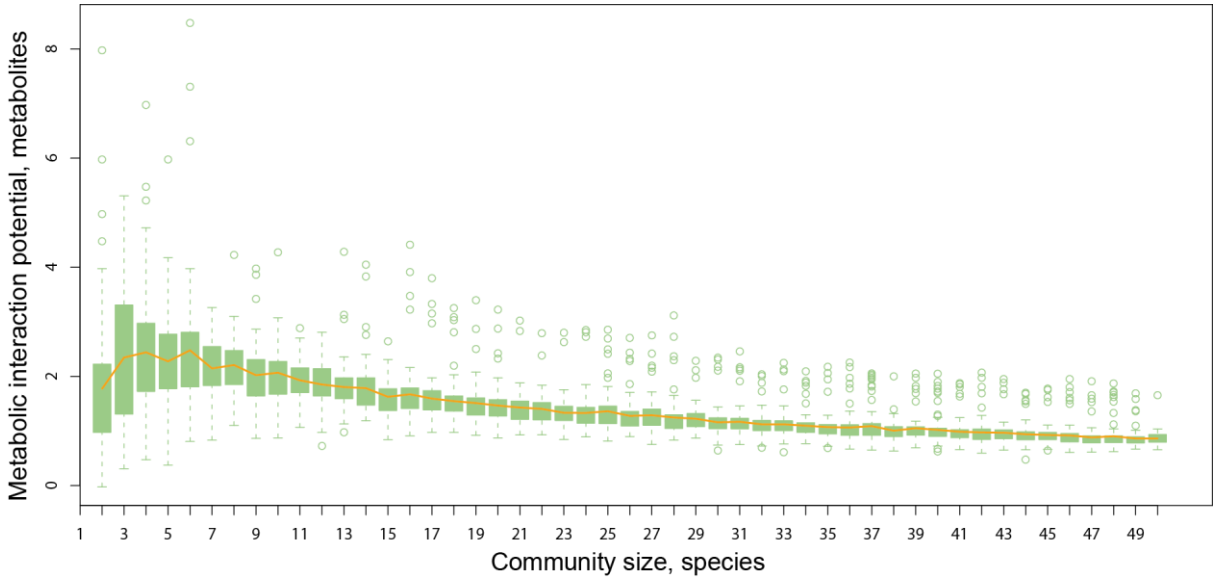


Figure 6.4 Metabolic interaction potential as a function of community sizes. On average highest metabolic interaction potential is observed within community of 6 species. Each boxplot represent a group of species of a given size randomly sampled (1000 times) from 1503 metabolic models space.

Mutualistic metabolite cross-feeding is prevalent in co-occurring lineages

To study in depth the mechanism behind metabolic interaction potential, we developed the species metabolite coupling analysis (SMETANA) framework which allows estimating all possible metabolic interdependencies between species within a community. Under defined environmental conditions and given that all species have to have a non-zero growth, method is able to determine whether one member's growth is dependent on (coupled to) another member of the community. Furthermore, the methodology can determine metabolites that must be exchanged to have non-zero growth, including all possible metabolite alternatives between all individuals in a community (Methods). The main difference between this methodology and other approaches (Freilich et al, 2011; Stolyar et al, 2007; Wintermute & Silver, 2010; Zomorrodi & Maranas, 2012) is that it does not assume any community objective function. Current microbial community constraint based steady-state modeling approaches rely on the main objective, which is often maximization of growth of the community under given constraints (Zomorrodi & Maranas, 2012). However, to our knowledge there is no experimental evidence studying multi-species community objectives. Moreover, very recently it was shown that species evolved in polycultures in a majority of cases have much slower growth phenotypes than species evolved in monocultures in the same environment (Lawrence et al, 2012). What is more, SMETANA enumerates all possible metabolic exchanges, in contrast to one-steady-state-solution approaches (Freilich et al, 2011; Zomorrodi & Maranas, 2012), providing a holistic view of interspecies metabolic exchanges. From a practical point of view, our methodology is capable of

simulating communities of large sizes (100 species models tested) in computationally affordable time, making it unique modeling platform available today.

The outcome of SMETANA is a score which signifies how important the metabolite interaction is between two individuals living in a group of species. For example, in a group of three members, in the case if one species needs a certain metabolite for growth and two other species can produce it, the interaction through this metabolite between the consumer and producer will be less important, than if there was only one member of community which could make this metabolite. A detailed explanation of scoring system and mathematical formulation can be found in the Methods section. For validation, we applied methodology to assess interactions in a well-characterized three-species synthrophic microbial system, and simulations were in agreement with experimental results (**Supplementary Figure 6.4**).

Given that we observed a strong potential of metabolic interaction in the co-occurring groups of species we, next, wanted to investigate whether particular type of interactions (commensal or mutualistic) were dominant in co-occurring communities and which of the metabolites were causing interactions under minimal medium conditions. The results from metabolite coupling analysis (**Supplementary Figure 6.1**) support our finding that co-occurring species tend to have high metabolic interaction potential. Metabolic coupling scores were significantly higher (P-value < 10^{-7}, Wilcoxon rank-sum test) in co-occurring communities with twice larger median in comparison to the random control suggesting that metabolite exchanges are highly important for bacterial coexistence. Moreover, mutualistic interactions were significantly (P-value < 10^{-7}, Hypergeometric test) in co-occurring lineages, suggesting that co-occurring species potentially can back each other up by compensating for the lack of nutrients and thus improving the overall fitness of individuals. Qualitative assessment of observed interactions showed that exchanges of amino-acids and carbohydrates were observed more frequently (P-value < 10^{-7}, Wilcoxon rank-sum test) in co-occurring bacterial species than in random combinations (**Figure 6.3**B).

Overall, the signals of metabolic interaction potential and metabolic coupling scores were substantially higher than differences of phylogenetic distances and metabolite resource overlaps in co-occurring communities (**Figure 6.3**C), signifying the importance of cooperation in microbial coexistence.

Discussion

The ecology of microbial communities is shaped by evolution and is reflected in variety of phenotypes in diverse microbiota. Because phenotype is the fundamental basis of interactions, and phenotypes are often specific to certain taxa, one should expect that the phylogenetic composition of a community is partially a product of species interactions. Among the oldest evolutionary hypotheses, proposed by Darwin in 1859 in *The Origin of Species*, the Naturalization hypothesis suggests that the conflict of existence is stronger between closely related species. Although phylogenetic distance and metabolic resource overlap were highly correlated (r = -0.48, P-value < 10^{-7}, **Figure 6.1**B), and both displayed statistically significant signals in co-occurring microbial lineages, neither of them were substantially different from random species combinations (**Supplementary Figure 6.1**). This result suggests that the competition for resources does not play the dominant role in the evolution of co-existing microbial species, otherwise we would expect the phylogenetic distance in co-occurring bacteria to be substantially larger. On the other hand, our proposed measure of species metabolic interaction potential shows remarkable statistically significant differences when compared to the random control (**Figure 6.3**A, **Supplementary Figure 6.1**C). Furthermore, under nutritionally minimal conditions, metabolite interactions in communities are more important than in random groups, as was indicated by high species metabolic coupling scores (**Supplementary Figure 6.1**D). More importantly, mutualistic interactions were dominant (P-value < 10^{-7}, Hypergeometric test) in co-occurring lineages, an indication that cooperation by metabolite cross-feeding potentially plays an important role in the shaping of microbial communities.

The analysis of present study is based on distributions of 16S rRNA sequences which were isolated from samples across a variety of natural environments. The sequences were grouped according to their similarities into operational taxonomic units (Chaffron et al, 2010), and a relatively small part (~11%) of the total dataset was mapped to the 16S gene of known sequenced bacterial genomes using our stringent criteria (Methods). Additionally, the majority of bacteria which are currently characterized and sequenced are mainly studied due to their medical and socio-economic interests. As a result, the OTU-to-genome mapping is potentially biased towards a particular group of organisms. Another potential source of bias can arise due to the numerous many-OTU-to-one-genome mappings unavoidably affecting the statistics of co-occurrence testing. To address this problem and partially overcome such biases, we removed combinations containing the most frequently appearing genomes (top 3 accounting for nearly 52% of all co-occurring quadruplets). The resulting trend of metabolic interaction potential in co-occurring lineages remained unchanged

(**Supplementary Figure 6.2**). As for metabolic reconstructions, we critically and carefully addressed issues which could potentially arise due to the automatic reconstruction process (Methods). All comparisons in this study were contrasted against random combinations of species. However, there exists the possibility that some groups of species would never be observed together in the same sample. To ensure that observed metabolic signals were capturing actual bacterial co-occurrence phenomenon and exclude possible biases that could be arisen due to randomization procedure, additionally, we repeated the random control procedure by choosing only combinations of species taken from the same samples. The alternative randomization procedure did not affect the overall results, signifying the importance of metabolic interactions in microbial co-existence (**Supplementary Figure 6.3**).

In conclusion, we have shown that metabolic interactions are prominently involved in microbial coexistence phenomena. Evolutionarily dissimilar microbial lineages did not form co-occurring communities more frequently than random, suggesting that competition is not the main determinant of microbial coexistence. A high number of mutualistic interactions in coexisting bacteria indicates that cooperation is important for microbial coexistence. Taken as a whole, it seems that the role of cooperation outweighs competition for resources in shaping microbial community structure.

Methods

Mapping OTU to genomes

16S rRNA sequence data clustered at 97% sequence identity and defined as operational taxonomic unit (OTU) was obtained from (Chaffron et al, 2010). Data represents distribution of OTUs in environmental samples, where each OTU representative sequence is mapped to a “sampling event”. Each “sampling event” is defined as the unique concatenation of three annotation fields: author, title and isolation source.

To map operational taxonomic units to the fully sequenced genomes, we extracted all 16S rRNA genes from Kyoto Encyclopedia of Genes and Genomes (KEGG), using the KEGG API (http://www.genome.jp/kegg/soap). We then used BLAST (Altschul et al, 1997) to compare obtained genome sequences against OUT representative sequences. The criteria for BLAST search were that at least 95% sequence identity and at least 95% of alignment of query sequence had to be matched and overlapped. In case if there was more than one 16S rRNA gene in genome, we used the longest sequence. Each OTU was mapped only once to a highest-ranking genome (highest bit score), but genome could be mapped to many OTUs (**Supplementary Table 6.1**).

Co-occurrence analysis

We tested co-occurrence significance for all possible pairs, combinations of threes and fours using Fisher's exact test. When testing co-occurrence for more than a pair, we have to consider more than one test. For example in case if we want to test whether species ABC are tend to co-occur more often than any of them alone, it is necessary to count the sites there BC are present but not A, AC are present but not B and AB but not C. Thus per triplet there should be tested 3 contingency tables, for combinations of fours correspondingly four.

To compute a large number of combinations in reasonable time we implemented a procedure which uses only *possible* combinations but not those which do not exist in any sample. For example, it would compute all combinations at each sampling site and skipping counting if such combination was already counted, thus combination s which do not exist at any sampling site will not be considered for counting. *P*-values obtained from multiple combinations tests were subsequently adjusted for false discovery (FDR) rate controlling procedure (Benjamini and Hochberg, 1995) as implemented in R Bioconductor (www.r-project.org, www.bioconductor.org) 'multtest' package.

Metabolic reconstructions and modeling

Using an automated genome-scale metabolic reconstruction pipeline Model SEED API (Henry et al, 2010) we reconstructed a total of 1503 unique simulation-ready bacterial models. Briefly, as input for reconstruction a genome sequence is provided, which is subsequently annotated and numerous standardized procedures (Feist et al, 2009) are applied to build simulation ready metabolic network. Metabolic model consists of complete set of enzymatic reactions, transport reactions and biomass equations, all reactions are represented in a stoichiometric matrix, written as a mass balances around metabolites, explained in details (Orth et al, 2010). A convention of representing multi-species model was borrowed from (Stolyar et al, 2007), except that now we extended it for more than two species model.

To avoid inconsistencies in automatic reconstructions, each model was additionally modified. Specifically, we noticed that reaction directions were inconsistent across the reconstructions. To compensate for these differences we extracted all irreversible reactions from manually reconstructed models (in Model SEED reaction space) and used it as an evidence for direction of irreversible reactions. In case if there were inconsistencies in manually reconstructions, we used directions which were in more than a half models. We formulated a mixed integer linear optimization problem to fix maximal number of irreversible reaction directions, while keeping possibility to produce biomass. Furthermore, by examining extracellular transporter reactions, we noticed that in

majority of cases models had dipeptides transport reactions but no individual transport for corresponding amino acids. Manual inspection of models showed that pathways for degradation of these amino acids existed in reconstructions. We concluded that such inconsistency could be an artifact of auto completion and gap-filling procedures implemented in automated reconstruction pipeline. To address this issue and avoid further false interpretations of results, for each dipeptide we also added corresponding two amino acid transporter reactions.

Metabolic interaction potential

As defined in main text, metabolic interaction potential is a difference in minimal number of components required for growth for a non-interacting community *vs.* an interacting community. Non-interacting community is represented by a multi-species model, except that each specie-compartment cannot excrete metabolite to a common environment, but metabolites are excreted from the multispecies-model. To find minimal number of metabolite to sustain community growth we implemented "Search for Minimal Media (SMM) Algorithm" described in (Klitgord & Segre, 2010). Metabolic interaction potential is computed using Equation 1.

$$MIP = \frac{M - I}{n} \qquad (1)$$

Here, *MIP* is metabolic interaction potential; *M* is the number of metabolites required for growth of non interacting community, correspondingly *I* is the number of metabolites required to support growth of interacting community and *n* is number of species in community.

Phylogenetic distance

Phylogenetic tree for 847 bacterial species using phylogenetic marker genes (Ciccarelli et al, 2006) and was kindly provided by Daniel Mende (Peer Bork Lab, EMBL). Phylogenetic distance for more than two species was computed as average distance between all combinations of pairs. For species which were not present in the tree, a random species was chosen from the same genus.

Metabolic resource overlap

To estimate metabolic resource overlap of species within community, we computed minimal media for whole non-interacting multi-species model. The obtained minimal media was used to constrain multi-species model (same model). Under such constraints we estimated minimal requirements of individual community member. Equation 2 was used to compute metabolic resource overlap,

$$MOP = \frac{N \sum_{i,j|i\neq j} M_i \cap M_j}{{}^{N}C_2 \sum_{i=0}^{N} M_i} \quad (2)$$

where *M* are minimal growth requirements for individual specie, N is the total number of members in community.

Species metabolic coupling score

Species metabolic coupling score reflects level of dependency of species *A* on metabolite *m* produced by species *B* under minimal media MM and is a product of SCS (Species Coupling Score), MUS (Metabolite Uptake Score) and MPS (Metabolite Production Score) scores which are explained later.

$$SMETANA = SCS \times MUS \times MPS$$

SMETANA score is distributed between 0 and 1 where values closer to zero reflect marginal dependency of species A on metabolite m produced by species B. On the other hand, scores closer to 1 denote more important links. As for boundary conditions SMETANA score of 0 means no relationship at all while SMETANA score of 1 means an essential dependency.

Species Coupling Score

Species Coupling Score (SCS) is a measure of dependence of species *A* on species *B* in community *C*. It is a fraction of cases where species *A* relies on products of donor *B* in direct or indirect manner over all cases. Implementation of SCS involves iteratively solving mixed integer linear programming problem, finding minimal set of donor species essential to support growth of species *A* in community *C*. Each time such set is found it is saved and a constraint eliminating it from further solution space is added to the program. The loop is repeated until there are no more such sets to be found.

```
Sc_scores = {}
for A in C:
  solutions = []
  do:
    species_set = minimize_donors_set(A, C, solutions)
    solutions += species_set
  while species_set != ∅
  for B in C:
    if A != B:
      sc_scores[(A, B)] = len([solutions containing B])/len(solutions)
```

As mentioned before `minimize_donors_set` routine solves mixed integer linear programming problem where objective is to minimize number of donating species in community *C* (1) while retaining growth of species *A* (4) and satisfying steady-state assumption (2). Other constraints include known or assumed flux bounds (3) and constraints controlling on/off state of other species in community (5,6). Later works by introducing binary variables θ_s (5). When θ_s is equal to 0 sum of all secretion fluxes is equal to 0. In other words, all secretion reactions are blocked in such case. Additionally, a constraint ensuring enabled species biomass production is added (6). Finally, to enumerate all possible solutions, each time a new solution is found a constraint blocking it from further search space is added (7).

$minimize\ \sum_{s\in C\setminus A}\theta_s$ (1)

$subject\ to$:

$S_s v_s = 0\ ,\ \forall s \in C$ (2)

$v^{lb} \leq v \leq v^{up}$ (3)

$v_{A,growth} = 1$ (4)

$\sum_{S\in L} v_{s,prod} - \gamma \cdot \theta_s \leq 0, \forall S \in C, \theta_s \in \{0,1\}, \gamma > argmax(v)$ (5)

$v_{A,growth} - v_{A,\min\_growth} * \theta_s \geq 0\ ,\quad \forall s \in C\setminus A$ (6)

$\sum_{s\in L}\theta_s < |\mathrm{L}|\ ,\ \forall L \in \{previously\ found\ solutions\}$ (7)

$-\varepsilon \leq v_{vit.uptake} - v_{measured\ vit.uptake} * v_{vit.uptake} \leq \varepsilon$

Metabolite Uptake Score

Metabolite Uptake Score (MUS) is a measure representing the extent to which species *A* relies on receiving metabolite *m* from other community members. In essence MUS is a fraction of minimal sets of received metabolites where metabolite *m* is present over all other minimal received sets. To calculate MUS first all alternative minimal sets of received metabolites for species *A* are found iteratively. Then a fraction of sets where metabolite *m* appears over all sets is calculated. This routine is repeated for each species in community *C*.

```
mu_scores = {}
for A in C:
```

```
solutions = []
do:
  metabolites_set = minimize_received_metabolites_set(A, C, solutions)
  solutions += metabolites_set
while metabolites_set != ∅
for m in A.received_metabolites:
  mu_scores[(A, m)] = len([solutions containing m])/len(solutions)
```

To find a minimal set of received metabolites `minimize_received_metabolites_set` routine solves mixed integer linear programming problem where objective is to minimize number active uptake reactions associated with received metabolites (1) while ensuring growth of species A (4) and satisfying steady-state assumption (2). Other constraints include known or assumed flux bounds (3) and constraints controlling on/off state of uptake fluxes (5). To enumerate all possible solutions, each time a new solution is found a constraint blocking it from further search space is added (6).

$$\textit{Minimize} \sum_{m \in \{metabolites\ from\ A\}} \theta_m \quad (1)$$

subject to:

$$S_s v_s = 0, \quad \forall s \in C \quad (2)$$

$$v^{lb} \le v \le v^{up} \quad (3)$$

$$V_{A,growth} = 1 \quad (4)$$

$$v_m - \gamma * \theta_m \le 0, \quad \forall m \in \{\text{metabolites uptakes from } A\} \quad (5)$$

$$\sum_{m \in L} \theta_m < |L|, \quad \forall L \in \{\text{previously found solutions}\} \quad (6)$$

$$-\varepsilon \le v_{vit.uptake} - v_{measured\ vit.uptake} * v_{vit.uptake} \le \varepsilon$$

Metabolite Production Score

Metabolite Production Score (MPS) is a binary value showing whether species *B* can produce metabolite m under given MM.

```
mp_scores = {}
```

```
for B in C:
  for m in {metabolites produced by B}
    mp_scores[(B, m)] = maximize_metabolite_yield(m, C) >= 1
```

To check whether species *B* can produce metabolite *m* in community *C* we iteratively solve linear program (`maximize_metabolite_yield`) where we try to maximize flux of secretion reaction of metabolite *m* in species *B* (1). As with other routines we assume steady-state of each organism separately (2) and constrain values of measured fluxes (3).

$$\textit{Maximize} \quad \sum_{m \in \{metabolites\ from\ A\}} v_m \qquad (1)$$

subject to:

$$S_s v_s = 0, \quad \forall s \in C \qquad (2)$$

$$v^{lb} \le v \le v^{up} \qquad (3)$$

$$-\varepsilon \le v_{vit.uptake} - v_{measured\,vit.uptake} * v_{vit.uptake} \le \varepsilon$$

Removing SEED model artifacts

While performing simulations we noticed that some organisms were growing by relying solely on vitamins which we attributed to the nature of automatic reconstruction of SEED models. We addressed this issue by introducing a blacklist of compounds whose use would be limited to their primary functions. Specifically, we calculated minimal uptakes of these compounds such that it would not damage organism growth. Having these minimal uptakes we added a constraints to all our routines ensuring selected compounds would be used only at their minimal rates ($\pm\varepsilon$ =0.01)

$$-\varepsilon \le v_{vit.uptake} - v_{measured\,vit.uptake} * v_{vit.uptake} \le \varepsilon$$

Statistical analysis

Pearson correlation coefficients and *P*-values for null hypothesis of no correlation (regression slope = 0) between average phylogenetic distance, metabolite resource overlap and metabolic interaction potential were calculated using statistical software R (www.r-project.org). For distribution comparisons we used Wilcoxon rank-sum test. Random distributions were generated by randomly choosing 10,000 times combinations of pairs, triplets and quadruplets from mapped genome space.

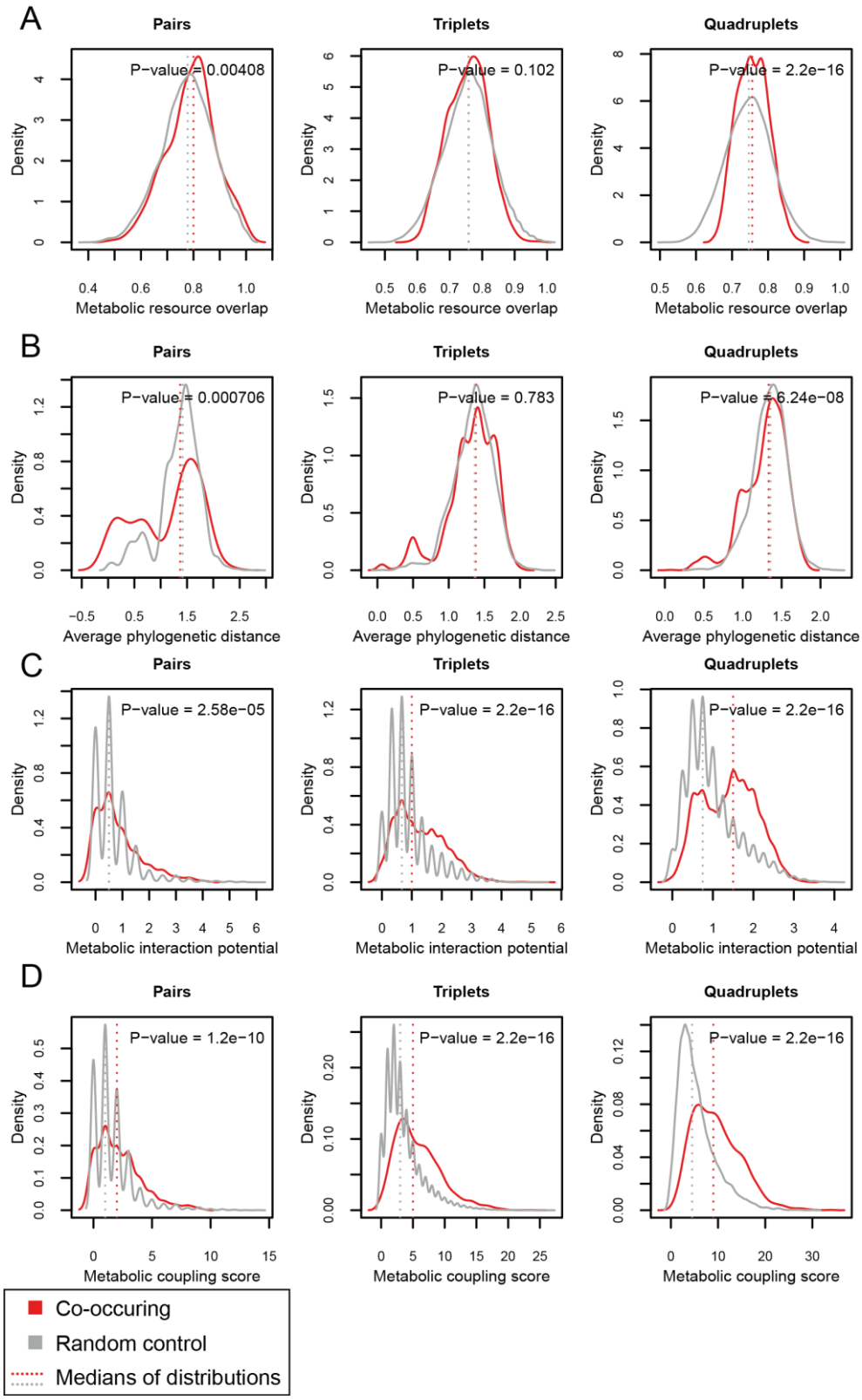
Supplementary Information

List of Supplementary Figures

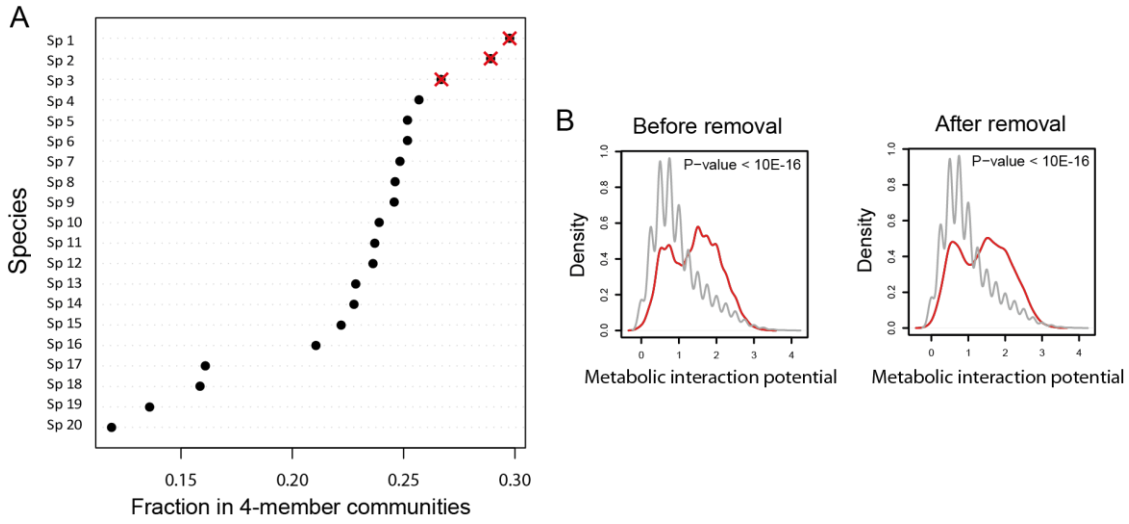
List of Supplementary Tables

Supplementary Table 6.1 Summary of identified communities and statistics of mappings.

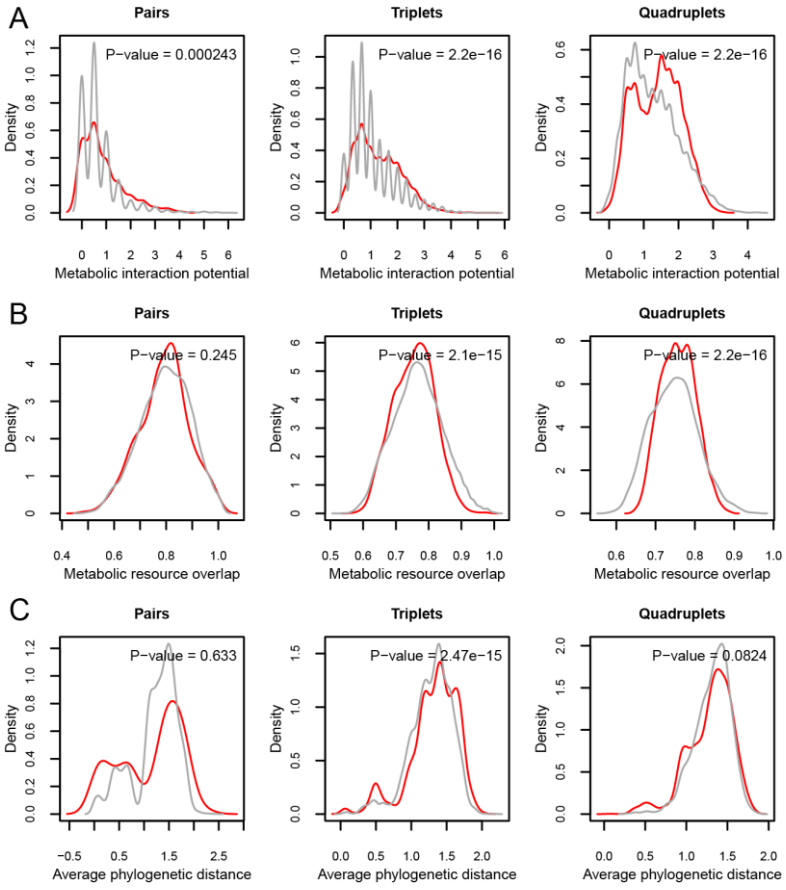
Number of OTUs with 97% sequence identity	5006			
Number of OTUs mapped with 95% sequence identity and 95% appearing at least 3 times among samples	536			
Number of mapped genomes	261			
Number sampling sites in which mapped genomes were present	1297			
	pairs	**triplets**	**quadruplets**	**total**
Co-occurring groups identified, FDR ≤ 0.01	381	3322	3518	7221
Number of possible communities observed in samples	2379	21570	77664	101613
Unique number of genomes/OTUs observed in significant pairs, FDR ≤ 0.01	127	129	49	150



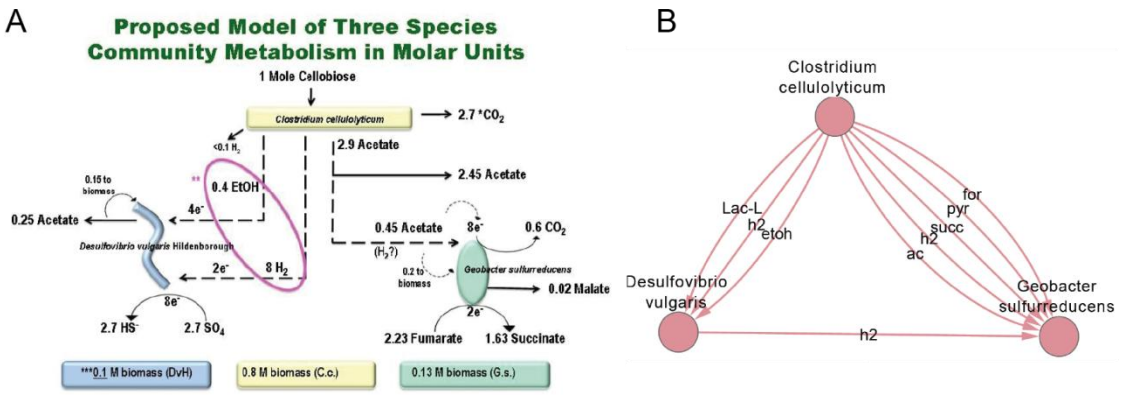
Supplementary Figure 6.1 Summary of all results comparing all metrics calculated in present study. Neither metabolic resource overlap (A) nor phylogenetic distances (B) in co-occurring communities were substantially different from random control. In contrast, both cooperation measures (C, D) show remarkable differences comparing to the random controls. Controls were computed by randomly sampling 10000 times species combinations of corresponding group sizes (grey curves). Importantly, random sampling was performed only from mapped genome space (261 metabolic reconstructions).



Supplementary Figure 6.2 Sensitivity analysis of results for metabolic interaction potential. After removal of top 3 species (A) found in 52% of all identified communities of 4 members, result remained unchanged (B). Results were valid for all co-occurring communities (data not shown).



Supplementary Figure 6.3 Summary of all results comparing all metrics calculated in present study using alternative random control. A) Alternative random control procedure did not change overall trend of metabolic interaction potential in co-occurring communities. Controls were computed by randomly sampling 10000 times (or all possible combinations) species combinations of corresponding group sizes (grey curves) only if species were present in the same habitat. Importantly, random sampling was performed only from mapped genome space (261 metabolic reconstructions). Neither metabolic resource overlap (B) nor phylogenetic distances (C) in co-occurring communities were substantially different from random control.



Supplementary Figure 6.4 Experimentally determined interspecies interactions (A) are in agreement with SMETANA predicted interactions. Our proposed method identified more possible interactions needed for three-species growth (B). Panel A adopted from (Miller et al, 2010).

Chapter 7

Conclusions and Future perspectives

The work presented in this thesis dealt with the development of methodologies and tools for understanding the principles behind the regulation and operation of metabolic networks. Emergent regulation of metabolic networks is one of the key messages of the presented work. It appears that a majority of metabolites in *Saccharomyces cerevisiae* at the transcriptional level are regulated in an organized manner by balancing the expression of neighboring enzymes. The principles described in **Chapter 3**, **Chapter 4** are solely dependent on network topology and possibly can be used to study transcriptional regulation in other organisms, including humans. The presented findings can be potentially applied to understand the pathogenesis of diseases in terms of perturbed metabolic phenotypes in conditions/diseases such as obesity, type 2 diabetes, cancer and many more. In particular, a straightforward application would be comparing regulation patterns in a healthy population and individuals with disease phenotypes. Hypothetically one would expect that regulation around certain metabolite nodes would change in a disease state. For example, co-regulation would become prevalent to specific pathways; consequently, one could stratify regulatory programs common to disease phenotypes or to certain population groups. Another possible application is to understand a short-term evolutionary adaptation of engineered strains. For instance, a common practice after performing strain genetic modifications is to make an adaptation experiment for the cell to adjust its metabolic phenotype to a more thermodynamically favorable state. Often this happens by means of genetic alterations or other unknown factors. I believe that such an adaptation leads to global changes of transcriptional regulation. If we would be able to understand how this phenomenon happens exactly, perhaps, in the future, we would be able to design and control strain adaptation without performing long-lasting experiments.

Another key finding presented in this work is that co-existing species have greater potential for metabolic cooperation than random control. Whether the observed trends of metabolic signal imply the causality (or at least degree) of microbial coexistence remains to be determined (if possible), nevertheless, the presented results suggest the importance of metabolic interactions in microbial coexistence. The work yielded the new tool (termed species metabolic coupling analysis) for modeling metabolic interactions and species interdependencies within microbial communities. The important feature of SMETANA is that it does not (and does not *need* to) assume any community objectives, making it unbiased and purely dependent on metabolic network stoichiometry. A straightforward application of the proposed methodology would be a design of artificial microbial

cultures for biodegradation purposes. For example, one might expect to observe steadiness/stability of communities if they were metabolically interdependent. Thus, under such constraints it is possible to design species combinations which could effectively degrade products of interest while keeping interaction among themselves. Furthermore, the species metabolic analysis has potential applications for elucidating host-pathogen interactions. It is known that some pathogens occupy certain body parts/organs of the human body. Determining pathogen dependencies on the host could potentially lead to discovery of novel drug targets.

The overall objectives of this thesis were to provide insights and develop methods for understanding operating principles of metabolic networks in microbial systems. Biological systems are very complex by their nature and achieving a complete, quantitative, and predictive understanding of system requires a close bridging between mathematical sciences and biology. One PhD project will not be able to answer all the questions, but I believe that the presented work brings the field one step towards understanding the operation of metabolism. As for me personally, the present work is just a beginning for future work directions.

References

Agarwal AK, Auchus RJ (2005) Minireview: cellular redox state regulates hydroxysteroid dehydrogenase activity and intracellular hormone potency. *Endocrinology* **146:** 2531-2538

Albert R (2005) Scale-free networks in cell biology. *Journal of cell science* **118:** 4947-4957

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25:** 3389-3402

Anderson RM, Latorre-Esteves M, Neves AR, Lavu S, Medvedik O, Taylor C, Howitz KT, Santos H, Sinclair DA (2003) Yeast life-span extension by calorie restriction is independent of NAD fluctuation. *Science* **302:** 2124-2126

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K et al (2011) Enterotypes of the human gut microbiome. *Nature* **473:** 174-180

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology* **2:** 2006 0008

Bajaj M, Defronzo RA (2003) Metabolic and molecular basis of insulin resistance. *Journal of nuclear cardiology : official publication of the American Society of Nuclear Cardiology* **10:** 311-323

Banez-Coronel M, Ramirez de Molina A, Rodriguez-Gonzalez A, Sarmentero J, Ramos MA, Garcia-Cabezas MA, Garcia-Oroz L, Lacal JC (2008) Choline kinase alpha depletion selectively kills tumoral cells. *Current cancer drug targets* **8:** 709-719

Bar-Even A, Flamholz A, Noor E, Milo R (2012) Rethinking glycolysis: on the biochemical logic of metabolic pathways. *Nature chemical biology* **8:** 509-517

Barabasi AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. *Nature reviews Genetics* **12:** 56-68

Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nature reviews Genetics* **5:** 101-113

Baxter CJ, Redestig H, Schauer N, Repsilber D, Patil KR, Nielsen J, Selbig J, Liu J, Fernie AR, Sweetlove LJ (2007) The metabolic response of heterotrophic Arabidopsis cells to oxidative stress. *Plant physiology* **143:** 312-325

Bennett BD, Kimball EH, Gao M, Osterhout R, Van Dien SJ, Rabinowitz JD (2009) Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nature chemical biology* **5:** 593-599

Boden G (1996) Fatty acids and insulin resistance. *Diabetes care* **19:** 394-395

Bradley PH, Brauer MJ, Rabinowitz JD, Troyanskaya OG (2009) Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae. *PLoS computational biology* **5:** e1000270

Braus GH (1991) Aromatic amino acid biosynthesis in the yeast Saccharomyces cerevisiae: a model system for the regulation of a eukaryotic biosynthetic pathway. *Microbiological reviews* **55:** 349-370

Brekasis D, Paget MS (2003) A novel sensor of NADH/NAD+ redox poise in Streptomyces coelicolor A3(2). *The EMBO journal* **22:** 4856-4865

Briggs GE, Haldane JB (1925) A Note on the Kinetics of Enzyme Action. *The Biochemical journal* **19:** 338-339

Brochado AR, Matos C, Moller BL, Hansen J, Mortensen UH, Patil KR (2010) Improved vanillin production in baker's yeast through in silico design. *Microbial cell factories* **9:** 84

Bruning T, Vamvakas S, Makropoulos V, Birner G (1998) Acute intoxication with trichloroethene: clinical symptoms, toxicokinetics, metabolism, and development of biochemical parameters for renal damage. *Toxicological sciences : an official journal of the Society of Toxicology* **41:** 157-165

Burgard AP, Nikolaev EV, Schilling CH, Maranas CD (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome research* **14:** 301-312

Burgard AP, Pharkya P, Maranas CD (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering* **84:** 647-657

Cakir T, Patil KR, Onsan Z, Ulgen KO, Kirdar B, Nielsen J (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. *Molecular systems biology* **2:** 50

Capel F, Klimcakova E, Viguerie N, Roussel B, Vitkova M, Kovacikova M, Polak J, Kovacova Z, Galitzky J, Maoret JJ, Hanacek J, Pers TH, Bouloumie A, Stich V, Langin D (2009) Macrophages and adipocytes in human obesity: adipose tissue gene expression and insulin sensitivity during calorie restriction and weight stabilization. *Diabetes* **58:** 1558-1567

Cavender-Bares J, Kozak KH, Fine PV, Kembel SW (2009) The merging of community ecology and phylogenetic biology. *Ecology letters* **12:** 693-715

Chaffron S, Rehrauer H, Pernthaler J, von Mering C (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome research* **20:** 947-959

Chen TH, Wang SY, Chen KN, Liu JR, Chen MJ (2009) Microbiological and chemical properties of kefir manufactured by entrapped microorganisms isolated from kefir grains. *Journal of dairy science* **92:** 3002-3013

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M et al (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research* **40:** D700-705

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* **311:** 1283-1287

Cimini D, Patil KR, Schiraldi C, Nielsen J (2009) Global transcriptional response of Saccharomyces cerevisiae to the deletion of SDH3. *BMC systems biology* **3:** 17

Cleland WW (1989) The kinetics of enzyme-catalyzed reactions with two or more substrates or products. I. Nomenclature and rate equations. 1963. *Biochimica et biophysica acta* **1000:** 213-220

Costenoble R, Picotti P, Reiter L, Stallmach R, Heinemann M, Sauer U, Aebersold R (2011) Comprehensive quantitative analysis of central carbon and amino-acid metabolism in Saccharomyces cerevisiae under multiple conditions by targeted proteomics. *Molecular systems biology* **7:** 464

Cox RP, Krauss MR, Balis ME, Dancis J (1974) Metabolic cooperation in cell culture: studies of the mechanisms of cell interaction. *Journal of cellular physiology* **84:** 237-252

Csete M, Doyle J (2004) Bow ties, metabolism and disease. *Trends in biotechnology* **22:** 446-450

Daran-Lapujade P, Rossell S, van Gulik WM, Luttik MA, de Groot MJ, Slijper M, Heck AJ, Daran JM, de Winde JH, Westerhoff HV, Pronk JT, Bakker BM (2007) The fluxes through glycolytic enzymes in Saccharomyces cerevisiae are predominantly regulated at posttranscriptional levels. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 15753-15758

David H, Hofmann G, Oliveira AP, Jarmer H, Nielsen J (2006) Metabolic network driven analysis of genome-wide transcription data from Aspergillus nidulans. *Genome biology* **7:** R108

del Sol A, Balling R, Hood L, Galas D (2010) Diseases as network perturbations. *Current opinion in biotechnology* **21:** 566-571

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 1777-1782

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO (2000) Protein function in the post-genomic era. *Nature* **405:** 823-826

Emmert-Streib F, Dehmer M (2011) Networks for systems biology: conceptual connection of data and function. *IET systems biology* **5:** 185-207

Falony G, Vlachou A, Verbrugghe K, De Vuyst L (2006) Cross-feeding between Bifidobacterium longum BB536 and acetate-converting, butyrate-producing colon bacteria during growth on oligofructose. *Applied and environmental microbiology* **72:** 7835-7841

Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. *Nature reviews Microbiology* **7:** 129-143

Fendt SM, Buescher JM, Rudroff F, Picotti P, Zamboni N, Sauer U (2010) Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. *Molecular systems biology* **6:** 356

Fleischman A, Kron M, Systrom DM, Hrovat M, Grinspoon SK (2009) Mitochondrial function and insulin resistance in overweight and normal-weight children. *The Journal of clinical endocrinology and metabolism* **94:** 4923-4930

Forster J, Famili I, Fu P, Palsson BO, Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome research* **13:** 244-253

Fraenkel DG (2011) Yeast Intermediary Metabolism. 1-38

Freilich S, Zarecki R, Eilam O, Segal ES, Henry CS, Kupiec M, Gophna U, Sharan R, Ruppin E (2011) Competitive and cooperative metabolic interactions in bacterial communities. *Nature communications* **2:** 589

Fuhrman JA (2009) Microbial community structure and its functional implications. *Nature* **459:** 193-199

Gallego O, Betts MJ, Gvozdenovic-Jeremic J, Maeda K, Matetzki C, Aguilar-Gurrieri C, Beltran-Alvarez P, Bonn S, Fernandez-Tornero C, Jensen LJ, Kuhn M, Trott J, Rybin V, Muller CW, Bork P, Kaksonen M, Russell RB, Gavin AC (2010) A systematic screen for protein-lipid interactions in Saccharomyces cerevisiae. *Molecular systems biology* **6:** 430

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20:** 307-315

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141-147

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G et al (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5:** R80

Gerosa L, Sauer U (2011) Regulation and control of metabolic fluxes in microbes. *Current opinion in biotechnology* **22:** 566-575

Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Molecular & cellular proteomics : MCP* **1:** 323-333

Hahn-Hagerdal B, Hallborn J, Jeppsson H, Meinander N, Walfridsson M, Ojamo H, Penttila M, Zimmermann FK (1996) Redox balances in recombinant Saccharomyces cerevisiae. *Annals of the New York Academy of Sciences* **782:** 286-296

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* **402:** C47-52

Haverkorn van Rijsewijk BR, Nanchen A, Nallet S, Kleijn RJ, Sauer U (2011) Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in Escherichia coli. *Molecular systems biology* **7:** 477

He J, Watkins S, Kelley DE (2001) Skeletal muscle lipid content and oxidative enzyme activity in relation to muscle fiber type in type 2 diabetes and obesity. *Diabetes* **50:** 817-823

Heinemann M, Sauer U (2010) Systems biology of microbial metabolism. *Current opinion in microbiology* **13:** 337-343

Henry CS, Broadbelt LJ, Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. *Biophysical journal* **92:** 1792-1805

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology* **28:** 977-982

Hentze MW, Preiss T (2010) The REM phase of gene regulation. *Trends in biochemical sciences* **35:** 423-426

Hirai MY, Klein M, Fujikawa Y, Yano M, Goodenowe DB, Yamazaki Y, Kanaya S, Nakamura Y, Kitayama M, Suzuki H, Sakurai N, Shibata D, Tokuhisa J, Reichelt M, Gershenzon J, Papenbrock J, Saito K (2005) Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. *The Journal of biological chemistry* **280:** 25590-25595

Holland WL, Brozinick JT, Wang LP, Hawkins ED, Sargent KM, Liu Y, Narra K, Hoehn KL, Knotts TA, Siesky A, Nelson DH, Karathanasis SK, Fontenot GK, Birnbaum MJ, Summers SA (2007) Inhibition of ceramide synthesis ameliorates glucocorticoid-, saturated-fat-, and obesity-induced insulin resistance. *Cell metabolism* **5:** 167-179

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292:** 929-934

Ihmels J, Levy R, Barkai N (2004) Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nature biotechnology* **22:** 86-92

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research* **31:** e15

Itani SI, Ruderman NB, Schmieder F, Boden G (2002) Lipid-induced insulin resistance in human muscle is associated with changes in diacylglycerol, protein kinase C, and IkappaB-alpha. *Diabetes* **51:** 2005-2011

Joshi-Tope G, Vastrik I, Gopinath GR, Matthews L, Schmidt E, Gillespie M, D'Eustachio P, Jassal B, Lewis S, Wu G, Birney E, Stein L (2003) The Genome Knowledgebase: a resource for biologists and bioinformaticists. *Cold Spring Harbor symposia on quantitative biology* **68:** 237-243

Kanehisa M (2002) The KEGG database. *Novartis Foundation symposium* **247:** 91-101; discussion 101-103, 119-128, 244-152

Karp PD, Paley S, Romero P (2002) The Pathway Tools software. *Bioinformatics* **18 Suppl 1:** S225-232

Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Jr., Assad-Garcia N, Glass JI, Covert MW (2012) A whole-cell computational model predicts phenotype from genotype. *Cell* **150:** 389-401

Kelley DE, He J, Menshikova EV, Ritov VB (2002) Dysfunction of mitochondria in human skeletal muscle in type 2 diabetes. *Diabetes* **51:** 2944-2950

Kersten S, Desvergne B, Wahli W (2000) Roles of PPARs in health and disease. *Nature* **405:** 421-424

Keurentjes JJ, Fu J, de Vos CH, Lommen A, Hall RD, Bino RJ, van der Plas LH, Jansen RC, Vreugdenhil D, Koornneef M (2006) The genetics of plant metabolism. *Nature genetics* **38:** 842-849

Kharchenko P, Church GM, Vitkup D (2005) Expression dynamics of a cellular metabolic network. *Molecular systems biology* **1:** 2005 0016

Klipp E, Nordlander B, Kruger R, Gennemark P, Hohmann S (2005) Integrative model of the response of yeast to osmotic shock. *Nature biotechnology* **23:** 975-982

Klitgord N, Segre D (2010) Environments that induce synthetic microbial ecosystems. *PLoS computational biology* **6:** e1001002

Klitgord N, Segre D (2011) Ecosystems biology of microbial metabolism. *Current opinion in biotechnology* **22:** 541-546

Koves TR, Li P, An J, Akimoto T, Slentz D, Ilkayeva O, Dohm GL, Yan Z, Newgard CB, Muoio DM (2005) Peroxisome proliferator-activated receptor-gamma co-activator 1alpha-mediated metabolic remodeling of skeletal myocytes mimics exercise training and reverses lipid-induced mitochondrial inefficiency. *The Journal of biological chemistry* **280:** 33588-33598

Kresnowati MT, van Winden WA, Almering MJ, ten Pierick A, Ras C, Knijnenburg TA, Daran-Lapujade P, Pronk JT, Heijnen JJ, Daran JM (2006) When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation. *Molecular systems biology* **2:** 49

Kumar A, Suthers PF, Maranas CD (2012) MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC bioinformatics* **13:** 6

Kummel A, Panke S, Heinemann M (2006) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Molecular systems biology* **2:** 2006 0034

Laughter AR, Dunn CS, Swanson CL, Howroyd P, Cattley RC, Corton JC (2004) Role of the peroxisome proliferator-activated receptor alpha (PPARalpha) in responses to trichloroethylene and metabolites, trichloroacetate and dichloroacetate in mouse liver. *Toxicology* **203:** 83-98

Lawrence D, Fiegna F, Behrends V, Bundy JG, Phillimore AB, Bell T, Barraclough TG (2012) Species interactions alter evolutionary responses to a novel environment. *PLoS biology* **10:** e1001330

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK et al (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298:** 799-804

Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature reviews Microbiology* **10:** 291-305

Li X, Gianoulis TA, Yip KY, Gerstein M, Snyder M (2010) Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses. *Cell* **143:** 639-650

Lin SJ, Defossez PA, Guarente L (2000) Requirement of NAD and SIR2 for life-span extension by calorie restriction in Saccharomyces cerevisiae. *Science* **289:** 2126-2128

Luttik MA, Vuralhan Z, Suir E, Braus GH, Pronk JT, Daran JM (2008) Alleviation of feedback inhibition in Saccharomyces cerevisiae aromatic amino acid biosynthesis: quantification of metabolic impact. *Metabolic engineering* **10:** 141-153

Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, Goryanin I (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular systems biology* **3:** 135

Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering* **5:** 264-276

Marstrand TT, Frellsen J, Moltke I, Thiim M, Valen E, Retelska D, Krogh A (2008) Asap: a framework for over-representation statistics for transcription factor binding sites. *PloS one* **3:** e1623

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S et al (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic acids research* **31:** 374-378

McKnight SL (2010) On getting there from here. *Science* **330:** 1338-1339

Metallo CM, Vander Heiden MG (2010) Metabolism strikes back: metabolic flux regulates cell signaling. *Genes & development* **24:** 2717-2722

Michaelis L, Menten M (1913) Die kinetik der invertinwirkung. *Biochem Z* **49**

Mikeskova H, Novotny C, Svobodova K (2012) Interspecific interactions in mixed microbial cultures in a biodegradation perspective. *Applied microbiology and biotechnology* **95:** 861-870

Miller LD, Mosher JJ, Venkateswaran A, Yang ZK, Palumbo AV, Phelps TJ, Podar M, Schadt CW, Keller M (2010) Establishment and metabolic analysis of a model microbial community for understanding trophic and electron accepting interactions of subsurface anaerobic environments. *BMC microbiology* **10:** 149

Moller S, Sternberg C, Andersen JB, Christensen BB, Ramos JL, Givskov M, Molin S (1998) In situ gene expression in mixed-culture biofilms: evidence of metabolic interactions between community members. *Applied and environmental microbiology* **64:** 721-732

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D et al (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* **34:** 267-273

Muoio DM, Newgard CB (2008) Mechanisms of disease: molecular and metabolic mechanisms of insulin resistance and beta-cell failure in type 2 diabetes. *Nature reviews Molecular cell biology* **9:** 193-205

Murray DB, Beckmann M, Kitano H (2007) Regulation of yeast oscillatory dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 2241-2246

Newgard CB, An J, Bain JR, Muehlbauer MJ, Stevens RD, Lien LF, Haqq AM, Shah SH, Arlotto M, Slentz CA, Rochon J, Gallup D, Ilkayeva O, Wenner BR, Yancy WS, Jr., Eisenson H, Musante G, Surwit RS, Millington DS, Butler MD et al (2009) A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell metabolism* **9:** 311-326

Nielsen J (2003) It is all about metabolic fluxes. *Journal of bacteriology* **185:** 7031-7035

Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Molecular systems biology* **5:** 320

Oliveira AP, Patil KR, Nielsen J (2008) Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. *BMC systems biology* **2:** 17

Oliveira AP, Sauer U (2012) The importance of post-translational modifications in regulating Saccharomyces cerevisiae metabolism. *FEMS yeast research* **12:** 104-117

Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nature biotechnology* **28:** 245-248

Patil KR (2006) Systems Biology of Metabolic Networks: Uncovering Regulatory and Stoichiometric Principles.

Patil KR, Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 2685-2689

Patti ME, Butte AJ, Crunkhorn S, Cusi K, Berria R, Kashyap S, Miyazaki Y, Kohane I, Costello M, Saccone R, Landaker EJ, Goldfine AB, Mun E, DeFronzo R, Finlayson J, Kahn CR, Mandarino LJ (2003) Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 8466-8471

Patti ME, Corvera S (2010) The role of mitochondria in the pathogenesis of type 2 diabetes. *Endocrine reviews* **31:** 364-395

Pehling G, Tessari P, Gerich JE, Haymond MW, Service FJ, Rizza RA (1984) Abnormal meal carbohydrate disposition in insulin-dependent diabetes. Relative contributions of endogenous glucose production and initial splanchnic uptake and effect of intensive insulin therapy. *The Journal of clinical investigation* **74:** 985-991

Periasamy S, Kolenbrander PE (2009) Aggregatibacter actinomycetemcomitans builds mutualistic biofilm communities with Fusobacterium nucleatum and Veillonella species in saliva. *Infection and immunity* **77:** 3542-3551

Petersen KF, Befroy D, Dufour S, Dziura J, Ariyan C, Rothman DL, DiPietro L, Cline GW, Shulman GI (2003) Mitochondrial dysfunction in the elderly: possible role in insulin resistance. *Science* **300:** 1140-1142

Petersen KF, Dufour S, Shulman GI (2005) Decreased insulin-stimulated ATP synthesis and phosphate transport in muscle of insulin-resistant offspring of type 2 diabetic parents. *PLoS medicine* **2:** e233

Phelan VV, Liu WT, Pogliano K, Dorrestein PC (2012) Microbial metabolic exchange--the chemotype-to-phenotype link. *Nature chemical biology* **8:** 26-35

Phielix E, Schrauwen-Hinderling VB, Mensink M, Lenaers E, Meex R, Hoeks J, Kooi ME, Moonen-Kornips E, Sels JP, Hesselink MK, Schrauwen P (2008) Lower intrinsic ADP-stimulated mitochondrial respiration underlies in vivo mitochondrial dysfunction in muscle of male type 2 diabetic patients. *Diabetes* **57:** 2943-2949

Poltak SR, Cooper VS (2011) Ecological succession in long-term experimentally evolved biofilms produces synergistic communities. *The ISME journal* **5:** 369-378

Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJ, Stern DF, De Virgilio C, Tyers M et al (2005) Global analysis of protein phosphorylation in yeast. *Nature* **438:** 679-684

Raghevendran V, Patil KR, Olsson L, Nielsen J (2006) Hap4 is not essential for activation of respiration at low specific growth rates in Saccharomyces cerevisiae. *The Journal of biological chemistry* **281:** 12308-12314

Rakoff-Nahoum S, Paglino J, Eslami-Varzaneh F, Edberg S, Medzhitov R (2004) Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell* **118:** 229-241

Ramirez de Molina A, Gallego-Ortega D, Sarmentero J, Banez-Coronel M, Martin-Cantalejo Y, Lacal JC (2005) Choline kinase is a novel oncogene that potentiates RhoA-induced carcinogenesis. *Cancer research* **65:** 5647-5653

Ramsey MM, Rumbaugh KP, Whiteley M (2011) Metabolite cross-feeding enhances virulence in a model polymicrobial infection. *PLoS pathogens* **7:** e1002012

Ray LB (2010) Metabolism. Metabolism is not boring. Introduction. *Science* **330:** 1337

Reaves ML, Rabinowitz JD (2011) Metabolomics in systems microbiology. *Current opinion in biotechnology* **22:** 17-25

Ristow M, Zarse K, Oberbach A, Kloting N, Birringer M, Kiehntopf M, Stumvoll M, Kahn CR, Bluher M (2009) Antioxidants prevent health-promoting effects of physical exercise in humans. *Proceedings of the National Academy of Sciences of the United States of America* **106:** 8665-8670

Roden M (2005) Muscle triglycerides and mitochondrial function: possible mechanisms for the development of type 2 diabetes. *International journal of obesity* **29 Suppl 2:** S111-115

Rodgers JT, Lerin C, Haas W, Gygi SP, Spiegelman BM, Puigserver P (2005) Nutrient control of glucose homeostasis through a complex of PGC-1alpha and SIRT1. *Nature* **434:** 113-118

Rossell S, van der Weijden CC, Kruckeberg AL, Bakker BM, Westerhoff HV (2005) Hierarchical and metabolic regulation of glucose influx in starved Saccharomyces cerevisiae. *FEMS yeast research* **5:** 611-619

Rossell S, van der Weijden CC, Lindenbergh A, van Tuijl A, Francke C, Bakker BM, Westerhoff HV (2006) Unraveling the complexity of flux regulation: a new method demonstrated for nutrient starvation in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* **103:** 2166-2171

Rutter J, Reick M, Wu LC, McKnight SL (2001) Regulation of clock and NPAS2 DNA binding by the redox state of NAD cofactors. *Science* **293:** 510-514

Salek RM, Maguire ML, Bentley E, Rubtsov DV, Hough T, Cheeseman M, Nunez D, Sweatman BC, Haselden JN, Cox RD, Connor SC, Griffin JL (2007) A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. *Physiological genomics* **29:** 99-108

Saltiel AR, Kahn CR (2001) Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **414:** 799-806

Sauer U, Zamboni N (2008) From biomarkers to integrated network responses. *Nature biotechnology* **26:** 1090-1092

Savage DB, Petersen KF, Shulman GI (2007) Disordered lipid metabolism and the pathogenesis of insulin resistance. *Physiological reviews* **87:** 507-520

Scarpulla RC (2006) Nuclear control of respiratory gene expression in mammalian cells. *Journal of cellular biochemistry* **97:** 673-683

Scarpulla RC (2008) Nuclear control of respiratory chain expression by nuclear respiratory factors and PGC-1-related coactivator. *Annals of the New York Academy of Sciences* **1147:** 321-334

Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, Kopka J, Willmitzer L, Zamir D, Fernie AR (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nature biotechnology* **24:** 447-454

Schellenberger J, Park JO, Conrad TM, Palsson BO (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics* **11:** 213

Schilling CH, Letscher D, Palsson BO (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of theoretical biology* **203:** 229-248

Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Molecular systems biology* **3:** 119

Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U (2012) Multidimensional optimality of microbial metabolism. *Science* **336:** 601-604

Schuster S, Dandekar T, Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in biotechnology* **17:** 53-60

Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature biotechnology* **18:** 326-332

Schuurmans JM, Rossell SL, van Tuijl A, Bakker BM, Hellingwerf KJ, Teixeira de Mattos MJ (2008) Effect of hxk2 deletion and HAP4 overexpression on fermentative capacity in Saccharomyces cerevisiae. *FEMS yeast research* **8:** 195-203

Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011) Global quantification of mammalian gene expression control. *Nature* **473:** 337-342

Seggewiss J, Becker K, Kotte O, Eisenacher M, Yazdi MR, Fischer A, McNamara P, Al Laham N, Proctor R, Peters G, Heinemann M, von Eiff C (2006) Reporter metabolite analysis of transcriptional profiles of a Staphylococcus aureus strain with normal phenotype and its isogenic hemB

mutant displaying the small-colony-variant phenotype. *Journal of bacteriology* **188:** 7765-7777

Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America* **99:** 15112-15117

Sen CK, Packer L (1996) Antioxidant and redox regulation of gene transcription. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **10:** 709-720

Shulman GI (2000) Cellular mechanisms of insulin resistance. *The Journal of clinical investigation* **106:** 171-176

Simpson RW, Shaw JE, Zimmet PZ (2003) The prevention of type 2 diabetes--lifestyle change or pharmacotherapy? A challenge for the 21st century. *Diabetes research and clinical practice* **59:** 165-180

Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3:** Article3

Sreekumar R, Halvatsiotis P, Schimke JC, Nair KS (2002) Gene expression profile in skeletal muscle of type 2 diabetes and the effect of insulin treatment. *Diabetes* **51:** 1913-1920

Stitt M, Sulpice R, Keurentjes J (2010) Metabolic networks: how to identify key components in the regulation of metabolism and growth. *Plant physiology* **152:** 428-444

Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, Stahl DA (2007) Metabolic modeling of a mutualistic microbial community. *Molecular systems biology* **3:** 92

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 9440-9445

Sweetlove LJ, Fell D, Fernie AR (2008) Getting to grips with the plant metabolic network. *The Biochemical journal* **409:** 27-41

Szallasi Z, Stelling J, Periwal V (2010) System Modeling in Cellular Biology: From Concepts to Nuts and Bolts. 464

Szendroedi J, Schmid AI, Chmelik M, Toth C, Brehm A, Krssak M, Nowotny P, Wolzt M, Waldhausl W, Roden M (2007) Muscle mitochondrial ATP synthesis and glucose transport/phosphorylation in type 2 diabetes. *PLoS medicine* **4:** e154

Tai SL, Boer VM, Daran-Lapujade P, Walsh MC, de Winde JH, Daran JM, Pronk JT (2005) Two-dimensional transcriptome analysis in chemostat cultures. Combinatorial effects of oxygen availability and macronutrient limitation in Saccharomyces cerevisiae. *The Journal of biological chemistry* **280:** 437-447

Terzer M, Stelling J (2008) Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics* **24:** 2229-2235

Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5:** 93-121

Trinh CT, Wlaschin A, Srienc F (2009) Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Applied microbiology and biotechnology* **81:** 813-826

Tu BP, Kudlicki A, Rowicka M, McKnight SL (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science* **310:** 1152-1158

Ueki K, Kondo T, Tseng YH, Kahn CR (2004) Central role of suppressors of cytokine signaling proteins in hepatic steatosis, insulin resistance, and the metabolic syndrome in the mouse. *Proceedings of the National Academy of Sciences of the United States of America* **101:** 10422-10427

Uemura H, Jigami Y (1992) Role of GCR2 in transcriptional activation of yeast glycolytic genes. *Molecular and cellular biology* **12:** 3834-3842

Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie AR (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO reports* **4:** 989-993

Usaite R, Nielsen J, Olsson L (2008a) Physiological characterization of glucose repression in the strains with SNF1 and SNF4 genes deleted. *Journal of biotechnology* **133:** 73-81

Usaite R, Wohlschlegel J, Venable JD, Park SK, Nielsen J, Olsson L, Yates Iii JR (2008b) Characterization of global yeast quantitative proteome data generated from the wild-type and glucose repression saccharomyces cerevisiae strains: the comparison of two quantitative methods. *Journal of proteome research* **7:** 266-275

Van Slyke DD, Cullen GE (1914) THE MODE OF ACTION OF UREASE AND OF ENZYMES IN GENERAL. *J Biol Chem* **19:** 141-180

Vander Heiden MG (2011) Targeting cancer metabolism: a therapeutic window opens. *Nature reviews Drug discovery* **10:** 671-684

Varma A, Palsson BO (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Applied and environmental microbiology* **60:** 3724-3731

Wang Q, Zhang Y, Yang C, Xiong H, Lin Y, Yao J, Li H, Xie L, Zhao W, Yao Y, Ning ZB, Zeng R, Xiong Y, Guan KL, Zhao S, Zhao GP (2010) Acetylation of metabolic enzymes coordinates carbon source utilization and metabolic flux. *Science* **327:** 1004-1007

Wang X, Gulbahce N, Yu H (2011) Network-based methods for human disease gene prediction. *Briefings in functional genomics* **10:** 280-293

Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, Deciu C, Winzeler E, Yates JR, 3rd (2003) Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 3107-3112

Westerhoff HV, Chen YD (1984) How do enzyme activities control metabolite concentrations? An additional theorem in the theory of metabolic control. *European journal of biochemistry / FEBS* **142:** 425-430

Wintermute EH, Silver PA (2010) Emergent cooperation in microbial metabolism. *Molecular systems biology* **6:** 407

Wisselink HW, Cipollina C, Oud B, Crimi B, Heijnen JJ, Pronk JT, van Maris AJ (2010) Metabolome, transcriptome and metabolic flux analysis of arabinose fermentation by engineered Saccharomyces cerevisiae. *Metabolic engineering* **12:** 537-551

Wolf YI, Karev G, Koonin EV (2002) Scale-free networks in biology: new insights into the fundamentals of evolution? *BioEssays : news and reviews in molecular, cellular and developmental biology* **24:** 105-109

Yang C, Hua Q, Shimizu K (2002a) Integration of the information from gene expression and metabolic fluxes for the analysis of the regulatory mechanisms in Synechocystis. *Applied microbiology and biotechnology* **58:** 813-822

Yang X, Pratley RE, Tokraks S, Bogardus C, Permana PA (2002b) Microarray profiling of skeletal muscle tissues from equally obese, non-diabetic insulin-sensitive and insulin-resistant Pima Indians. *Diabetologia* **45:** 1584-1593

Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T (2010) Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* **26:** i255-260

Ze X, Duncan SH, Louis P, Flint HJ (2012) Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *The ISME journal* **6:** 1535-1543

Zelezniak A, Pers TH, Soares S, Patti ME, Patil KR (2010) Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes. *PLoS computational biology* **6:** e1000729

Zhang Q, Piston DW, Goodman RH (2002) Regulation of corepressor function by nuclear NADH. *Science* **295:** 1895-1897

Zimmet P, Alberti KG, Shaw J (2001) Global and societal implications of the diabetes epidemic. *Nature* **414:** 782-787

Zomorrodi AR, Maranas CD (2012) OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS computational biology* **8:** e1002363